Princess Sumaya University for Technology
جامعة الأميرة سميّة للتكنولوجيا

صندوق دعم البحث العلمي والابتكار
Scientific Research and Innovation Support Fund

Jordanian Journal of Computers and Information Technology

September 2026 VOLUME 12 NUMBER 03
ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

JJCIT

PAGES	PAPERS
266 - 282	TOWARDS RELIABLE IOT AND WEB API MASHUPS: A HISTORY-AWARE LINKED DATA COMPOSITION APPROACH Mohammed Amine Belfedhal, Mimoun Malki and Abderrahmane Khiat
283- 296	ANALYTICAL INSIGHTS INTO THE AVERAGE BLOCK ERROR RATE OF FLUID ANTENNA SYSTEMS UNDER SHORT-PACKET CONSTRAINTS Chi-Bao Le and Tin Phu Tran
297 - 316	ADVANCEMENTS IN AI-DRIVEN DIABETIC RETINOPATHY DETECTION: A SYSTEMATIC SURVEY DIRECTIONS Yasser A. Ali, Ahmed Z. Emam, Muhammed H. Zayyan, Samir Elmougy and Ahmed S. Tolba
317 - 333	DUAL-LEVEL PROTOTYPE ALIGNMENT VIA CROSS- ATTENTION FOR FEW-SHOT REMOTE SENSING SEMANTIC SEGMENTATION Mustafa Alawadi and Mansoor Fateh
334 - 359	FR-CNN: SOFTWARE FAULT PREDICTION USING A NOVEL HYBRID HIERARCHICAL MODEL Mehrasa M. Jouybari, Alireza Tajary, Mansoor Fateh and Esmaeel Tahanian
360 - 387	NETWORK PERFORMANCE OF TWO-WAY RELAYING WITH FOUNTAIN CODING IN WIRELESS SENSOR NETWORKS UNDER CO-CHANNEL INTERFERENCE Duy-Hung Ha, Kien-Duc Tran, Dac-Binh Ha, Hien-Luong Nguyen, The-Anh Ngo and Quoc-Hao Truong
388 - 411	BREAST LESION CLASSIFICATION: LEAKAGE-FREE PATCH-BASED TRANSFER LEARNING WITH CROSS-DATASET EVALUATION ON CBIS-DDSM, MIAS AND INBREAST Hitesh Kag, Dinesh Kumar and Anjali Diwan
412 - 428	FILTER SPECTRAL EFFICIENCY (FSE): STRUCTUREDNETWORK PRUNING VIA INTER-LAYER SPECTRAL COUPLING Soumia Laroui, Hamida Seba, Karima Amrouche and Yassine Ait Ali Yahia

www.jjcit.org jjcit@psut.edu.jo

An International Peer-Reviewed Scientific Journal Financed by the Scientific Research and Innovation Support Fund

Jordanian Journal of Computers and Information Technology (JJCIT)

The Jordanian Journal of Computers and Information Technology (JJCIT) is an international journal that publishes original, high-quality and cutting edge research papers on all aspects and technologies in ICT fields.

JJCIT is hosted and published by Princess Sumaya University for Technology (PSUT) and supported by the Scientific Research Support Fund in Jordan. Researchers have the right to read, print, distribute, search, download, copy or link to the full text of articles. JJCIT permits reproduction as long as the source is acknowledged.

AIMS AND SCOPE

The JJCIT aims to publish the most current developments in the form of original articles as well as review articles in all areas of Telecommunications, Computer Engineering and Information Technology and make them available to researchers worldwide. The JJCIT focuses on topics including, but not limited to: Computer Engineering & Communication Networks, Computer Science & Information Systems and Information Technology and Applications.

INDEXING

JJCIT is indexed in:

EDITORIAL BOARD SUPPORT TEAM

LANGUAGE EDITOR

Haydar Al-Momani

EDITORIAL BOARD SECRETARY

Eyad Al-Kouz

All articles in this issue are open access articles distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

JJCIT ADDRESS

WEBSITE: www.jjcit.org
EMAIL: jjcit@psut.edu.jo
ADDRESS: Princess Sumaya University for Technology, Khalil Saket Street, Al-Jubaiha
B.O. BOX: 1438 Amman 11941 Jordan
TELEPHONE: +962-6-5359949
FAX: +962-6-7295534

EDITORIAL BOARD

Wejdan Abu Elhaija (EIC) — Ahmad Hiasat (Senior Editor)

Aboul Ella Hassanien — Adil Alpkoçak — Adnan Gutub

Adnan Shaout — Christian Boitet — Gian Carlo Cardarilli

Omer Rana — Mohammad Azzeh — Omar A. Alzubi

Ahmed Al-Taani — Lutfi Al-Sharif — Omar S. Al-Kadi

Raed A. Shatnawi — João L. M. P. Monteiro — Leonel Sousa

Omar Al-Jarrah — Angelo Lorusso

INTERNATIONAL ADVISORY BOARD

Ahmed Yassin Al-Dubai
UK

Albert Y. Zomaya
AUSTRALIA

Chip Hong Chang
SINGAPORE

Izzat Darwazeh
UK

Dia Abu Al Nadi
JORDAN

George Ghinea
UK

Hoda Abdel-Aty Zohdy
USA

Saleh Oqeili
JORDAN

João Barroso
PORTUGAL

Karem Sakallah
USA

Khaled Assaleh
UAE

Laurent-Stephane Didier
FRANCE

Lewis Mackenzies
UK

Zoubir Hamici
JORDAN

Korhan Cengiz
TURKEY

Marco Winzker
GERMANY

Marwan M. Krunz
USA

Mohammad Belal Al Zoubi
JORDAN

Michael Ullman
USA

Ali Shatnawi
JORDAN

Mohammed Benaissa
UK

Basel Mahafzah
JORDAN

Nadim Obaid
JORDAN

Nazim Madhavji
CANADA

Ahmad Al Shamali
JORDAN

Othman Khalifa
MALAYSIA

Shahrul Azman Mohd Noah
MALAYSIA

Shambhu J. Upadhyaya
USA

"Opinions or views expressed in papers published in this journal are those of the author(s) and do not necessarily reflect those of the Editorial Board, the host university or the policy of the Scientific Research Support Fund".

"ما ورد في هذه المجلة يعبر عن آراء الباحثين ولا يعكس بالضرورة آراء هيئة التحرير أو الجامعة أو سياسة صندوق دعم البحث العلمي والابتكار".

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

TOWARDS RELIABLE IOT AND WEB API MASHUPS: A HISTORY-AWARE LINKED DATA COMPOSITION APPROACH

Mohammed Amine Belfedhal[1], Mimoun Malki[2] and Abderrahmane Khiat[3]

(Received: 29-Jan.-2026, Revised: 23-Apr-2026, 25-May-2026 and 14-Jun.-2026, Accepted: 16-Jun.-2026)

ABSTRACT

The growing number of heterogeneous Web and IoT APIs poses a fundamental challenge to the reliability and reusability of modern applications, as existing API composition frameworks are hindered by semantic fragmentation and a neglect of historical usage patterns. To address this problem, this research paper introduces a history-aware semantic framework that enhances the long-term robustness of service compositions by embedding APIs into a unified knowledge graph via a dedicated ontology that harmonizes heterogeneous data and services through semantic abstraction and domain-level alignment. Reliability of composition is further enhanced by incorporating historical composition data capturing co-occurrence and popularity. A genetic algorithm dynamically optimizes the weights of these metrics to select the Top-K most reliable mashup compositions. Extensive evaluation on the large-scale ProgrammableWeb benchmark (11,492 APIs and 7,336 mashups) demonstrates a statistically significant improvement of 26% in mashup reliability ($p<0.01$) compared to state-of-the-art baselines. While this work focuses on Intelligent Media mashups as a representative application domain, the proposed framework is applicable to general Web and IoT API composition. The integration of semantic abstraction, historical analysis and optimization provides a robust solution to generate resilient, adaptive and highly reliable service mashups, offering a useful approach to support the long-term adaptability and reuse of complex service ecosystems.

KEYWORDS

Intelligent media, Media mashups, Semantic integration, IoT and Web APIs.

1. INTRODUCTION

Modern software applications increasingly rely on the composition of diverse Web and IoT services in order to deliver rich and adaptive functionalities. At the same time, there is a growing need for service compositions that are not only functional, but also robust and reusable over time. In this context, the rapid expansion of the Web [1] and the Internet of Things (IoT) [2] has created unprecedented opportunities, producing massive volumes of interconnected data from APIs, sensors, devices and services. However, this dynamism introduces significant challenges-namely, semantic heterogeneity, interoperability breakdowns and an inability to generate reliably reusable mashups [3].

Mashups have emerged as a practical means of harnessing this rich ecosystem by enabling the creation of custom applications that combine functionalities from multiple Web and IoT APIs [4], allowing for new business opportunities [5]. For example, geospatial APIs can be merged with social-media streams to create context-aware cultural or tourism applications.

A fundamental challenge in developing such applications lies in the composition of mashups, in which diverse data and service streams must be integrated into meaningful and cohesive services. Composition is the key step that enables heterogeneous Web and IoT APIs to be integrated into value-added applications [6]. In this paper, we define a mashup composition as the process of selecting and combining a set of APIs to fulfill a specific functional requirement, with an emphasis on reliability and reusability of the resulting service bundle. Yet, the richness of available APIs-from sensors and social platforms to mapping and multimedia services-also carries drawbacks. One major drawback is interoperability, due to APIs exposing information in incompatible formats, it relies on different communication protocols, adopting varying structural designs [7]. This heterogeneity stems from the multitude of providers, each with its own way of developing, publishing and delivering resources, This leads to interoperability breakdowns and integration issues that hinder adaptive and reliable services [8].

1. M. A. Belfedhal is with EEDIS Research Lab., Djillali Liabes Uni. of Sidi Bel Abbes, Algeria. Email: mohammed.belfedhal@gmail.com
2. M. Malki is with LabRI Research Laboratory, Ecole Supérieure en Informatique de Sidi Bel Abbes, Algeria. Email: m.malki@esi-sba.dz
3. A. Khiat is with EEDIS Research Lab. and Dr. Moulay Tahar University of Saida, Algeria. Email: khiat.abderrahmane@univ-saida.dz

One solution to overcome this heterogeneity is to apply Linked Data principles [7] which extend traditional mashups by employing a unified data model known as the Resource Description Framework (RDF) model to lift API descriptions and then combine them leading to the emergence of Semantic Mashup.

A review of current semantic mashup and API composition frameworks reveals persistent limitations that reduce their efficiency for building reliable and reusable applications. These include: (1) semantic fragmentation, where a lack of well semantically structured RDF-based models prevents full interoperability; (2) ignoring historical usage patterns, such as co-occurrence and popularity, which prevents the application from learning from through compositions; (3) fixed weighting, where reliance on unchanging parameters limits adaptability to evolving contexts; and (4) functional gaps, where recommendation methods focus on clustering or complementarity, but overlook the functional variety needed for diverse services.

In this research, we introduce a history-aware semantic approach that overcomes these gaps and advances the state of reliable service composition. Our approach transforms heterogeneous Web and IoT APIs, together with their composition history, into semantically enriched, well-structured knowledge representations. More precisely, by using Linked Data principles, APIs are lifted into RDF through a dedicated API ontology and aligned with relevant domain ontologies selected according to user requirements. This semantic enrichment not only enables the seamless integration of disparate service resources, but also supports the discovery of more accurate and contextually relevant compositions through semantic interoperability. Besides semantic similarity, our framework enhances reliability further by incorporating historical composition data, such as Co-Occurrence-the frequency with which APIs have been combined - and Popularity - the extent of their reuse and evaluation within existing mashups. These factors are dynamically optimized through a genetic algorithm to compute the best-performing compositions. By unifying APIs into a shared semantic space and embedding history-aware reliability analysis, the framework delivers more robust, adaptive and reusable integrations, thereby contributing to the development of next-generation service-oriented applications and services.

Sustainability encompasses environmental, economic, social and technical dimensions. In this work, we address only one aspect: reusability - the ability of a mashup to remain robust and adaptable over time, reducing maintenance effort. We do not claim to address other dimensions, such as energy consumption. Additionally, we evaluate our general framework on Intelligent Media mashups as a challenging representative domain (involving video, audio, social and IoT APIs) without implying coverage of all mashup types.

To summarize, the main contributions of this research work consist of:

- Designing and developing a new API ontology and a JSON-to-RDF mapping mechanism for the semantic enrichment of heterogeneous, service-oriented APIs.
- Introducing a history-aware reliability model that integrates functional similarity, co-occurrence and popularity metrics within the RDF-based descriptions.
- Developing an adaptive optimization mechanism using a genetic algorithm to dynamically balance these metrics for recommending the Top-K reliable service mashups.

The paper is structured as follows: Section 2 reviews related work and clearly delineates the research gaps. Section 3 details our proposed framework. Section 4 presents the experimental evaluation and discusses the results. Finally, Section 5 concludes the paper and outlines future work.

2. RELATED WORK

Efforts to enhance automation in service integration and composition have inspired numerous approaches that are increasingly relevant to modern distributed systems. By enriching Web and IoT services with semantic descriptions, these methods aim to facilitate adaptive and context-aware applications that can integrate heterogeneous resources.

Early foundational work [9] provided ontologies and rule-based models that laid the foundation for semantically enriched services. WSMO-Lite [10] extended this by semantically annotating RESTful services, enabling applications to link APIs with contextual data. Similarly, scripting approaches, such as "S" [11], defined interaction resources for dynamic integration, while behavior modeling tools, like

ReLL [12] and metadata exposure methods, such as LRDD [13], supported richer interaction with Web APIs. RESTdesc [14] and RESTdoc [15] applied RDF-based templates to facilitate semantic discovery and the W3C HTTP vocabulary in RDF [16] helped bridge Web communication with semantic interoperability. More formal interaction models, such as pi-calculus [17] and triple space computing [18], define rigorous service interaction frameworks. In parallel, Linked Web API datasets [19]-[20] and semantic link discovery approaches [21] contributed to organizing and connecting API knowledge at scale. User-centered efforts, such as SAREST [22], SAWADL [23], introduced semantic annotation standards to enable discovery and generation of mashups, while visual tools [24] allowed non-experts to assemble services *via* drag-and-drop interfaces. Collectively, these approaches established important foundations by showing how semantic enrichment can transform fragmented APIs into interoperable and reusable services.

IoT-driven systems have adopted mashup and semantic composition techniques to deliver adaptive, user-centric services. For instance, Node-RED [25] provides a drag-and-drop environment that helps creators prototype IoT-enabled workflows. SSC [26] composes applications using live Linked Data and SPARQL, enabling dynamic linking of sensor data with external sources. Standards, such as OneM2M [27], use semantic reasoning to match heterogeneous sensor data, facilitating rich, real-time integrations. To improve interaction and accessibility, tools, like CoMMA [28], assist natural-language customization of mashups, while visual-composition systems automate the assembly of services. Additionally, the use of aspect-oriented programming (AOP) to modularize cross-cutting concerns in cloud-based systems, as demonstrated by the WeaveSim framework [29], illustrates how separating adaptive behaviors from core logic enhances scalability and reusability. On the front of content recommendation, neural embeddings and matrix factorization are used to suggest service compositions; user behavior pattern analysis refines recommendations toward more personalized delivery [30]-[31].

Systems, such as The Art of Video MashUp [32], combine semantic annotations and media features to help non-expert users craft creative mashups and A New Social Media Mashup Approach [33] merges social feeds and event metadata to detect and present content based on time, topic and location. Despite these advances, few systems jointly address semantic enrichment, adaptive delivery, composition history and reusability. Similarly, [34] enhances content recommendation accuracy by analyzing user-behavior patterns, while SSR [35] addresses redundancy through functional ranking, CSR [36] mitigates overlaps *via* category-aware clustering and SPR [37] optimizes complementarity using topic/tag analysis and Quadratic Knapsack models.

Despite their contributions, existing approaches reveal critical limitations when applied to building reliable and reusable systems. First, semantic fragmentation persists, which is observed in frameworks, such as Node-RED [25] and SA-REST [22]. Those frameworks lack well-structured RDF-based models to infer relationships between heterogeneous services, leading to fragmented integration. Second, historical blindness is common in works, such as the ML framework proposed by Xu et al. [31] and STRIM [38], leading to the neglect of patterns, such as co-occurrence and popularity, thus preventing systems from learning from past compositions and limiting adaptive optimization. Third, static weighting constrains adaptability, since context-aware systems (e.g., Boulakbech et al. [30]) rely on fixed parameters rather than dynamically adjusting reliability scores in response to evolving contexts. Finally, functional gaps or oversight limit personalization, as recommendation techniques, like CSR [36] and SPR [37], prioritize clustering or complementarity, but overlook nuanced functional diversity needed for adaptive and personalized services.

To overcome these gaps, we propose a history-aware semantic framework that integrates Linked Data principles (*via* RDF-based semantic lifting) with composition-history analysis (co-occurrence and popularity) and semantic functional analysis to identify reusable patterns across services. Reliability, metrics-semantic functional similarity, historical co-occurrence and popularity-are dynamically balanced using a genetic algorithm, ensuring contextually optimized Top-K compositions. The novelty of our approach can be understood from three dimensions. First, the dedicated API ontology was developed specifically for this framework. Second, while the genetic algorithm is an established technique, its application to per-request weight tuning for mashup-reliability scoring is a novel adaptation. Third, the API ontology enables semantic enrichment and interoperability, while the genetic algorithm dynamically adjusts the reliability weights; their integration into a single pipeline yields a robust Top-K mashup recommendation that neither component could deliver alone. In doing so, our approach delivers semantically enriched, adaptive and reusable integrations that advance the state of

service composition by enabling more reliable, context-aware and maintainable applications.

Table 1. Comparison of service composition frameworks.

Framework	Data Representation	Semantic Alignment & Matching	Historical Metrics	Optimization / Adaptivity	Reusability & Reliability
SA-REST [22]	XML + OWL annotations	Partial (ontology-based only)	Not considered	None (static)	Low - limited adaptability
Node-RED [25]	JSON flow-based	No formal semantics	Not considered	Manual selection	Low - manual integration
SSR [34]	RDF triples	Semantic category ranking	No historical data	Static weighting	Moderate-redundancy
CSR [35]	Category ontology	Cluster-based similarity	Popularity only	Fixed parameters	Moderate - no adaptive learning
SPR [36]	Topic graph	Shallow semantic grounding	Popularity + topic correlation	Static scoring	Moderate - no semantic matching
Our framework	RDF + Linked Data ontology	Full semantic alignment + ontology mapping	Popularity and co-occurrence in RDF	Dynamic weight optimization (GA)	High - adaptive, reliable, reusable

3. TOP-K MEDIA-INTELLIGENT MASHUP COMPOSITION *VIA* LINKED DATA AND HISTORICAL USAGE

In this section, we describe our new approach-the History-aware Semantic Mashup approach-which aims to enable reliable, reusable and adaptive composition of Web and IoT services. The framework addresses the limitations of existing mashup systems - such as semantic fragmentation, static weighting and lack of historical awareness-by integrating semantic enrichment, composition-history analysis and adaptive optimization (dynamic weights) in a unified model. Although we illustrate the approach using media-centric examples and evaluate it on media mashup datasets, the methodology is applicable to any domain requiring reliable API composition. The novelty of the proposed approach lies in three complementary aspects, each contributing to the overall reliability and reusability of mashup generation: (1) API Ontology and JSON-to-RDF Mapping: Unlike previous frameworks that rely on generic lightweight RDF transformation, our approach introduces a specialized ontology tailored for Web and IoT APIs. It enables JSON-to-RDF transformation, providing uniform semantic representation of heterogeneous services and improving interoperability across API ecosystems. (2) History-Aware Reliability Modeling: The framework captures historical composition knowledge by analyzing two key metrics-popularity (aggregated user ratings and satisfaction scores) and co-occurrence (joint appearance of APIs in successful compositions). These metrics are stored as RDF properties, allowing the system to infer reliability based on real usage trends rather than static design assumptions. (3) Adaptive Optimization *via* Genetic Algorithm: To dynamically balance semantic similarity and historical reliability, our proposed solution employs a genetic algorithm (GA) that optimizes the weighting of reliability factors. The GA iteratively evolves reliability scores through crossover and mutation operations, ensuring that the selected Top-K mashups achieve the best trade-off between functionality, popularity and co-occurrence for each specific composition request. By unifying these three dimensions-semantic enrichment, historical awareness and adaptive optimization - the proposed framework overcomes the rigidity of existing mashup approaches and establishes a robust foundation for reliable service integration.

As shown in Figure 1, our contribution lies in three dimensions: (1) Semantic enrichment: API descriptions are lifted into RDF using our API ontology and an Ontology-based JSON-to-RDF Mapping technique [39], enabling APIs to be treated as semantically linkable with other resources. Once lifted, they are aligned with domain ontologies according to user requirements, ensuring interoperability across heterogeneous streams. (2) History-aware reliability: We incorporate composition history by modeling popularity (how widely APIs or mashups have been reused), co-occurrence (how frequently APIs appear

together in compositions, reflecting their compatibility) and functional similarity (semantic proximity of their capabilities). These factors reflect the experiential knowledge of past mashups, analogous to how user feedback and usage data guide recommendations. (3) Adaptive optimization: A genetic algorithm dynamically assigns optimal weights to these measures, ensuring that Top-K reliable mashups are selected per user request.

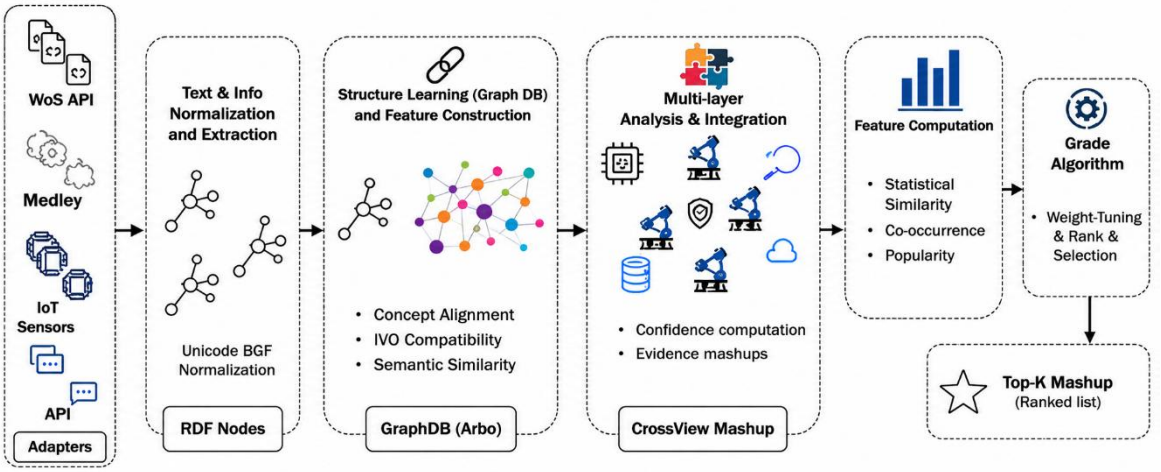


Figure 1. Selection of best API combinations.

We first present the overall architecture of our mashup framework; then, we highlight our main contributions.

3.1 Building Blocks of the Mashup Architecture

As shown in Figure 2, the framework is structured into five layers:

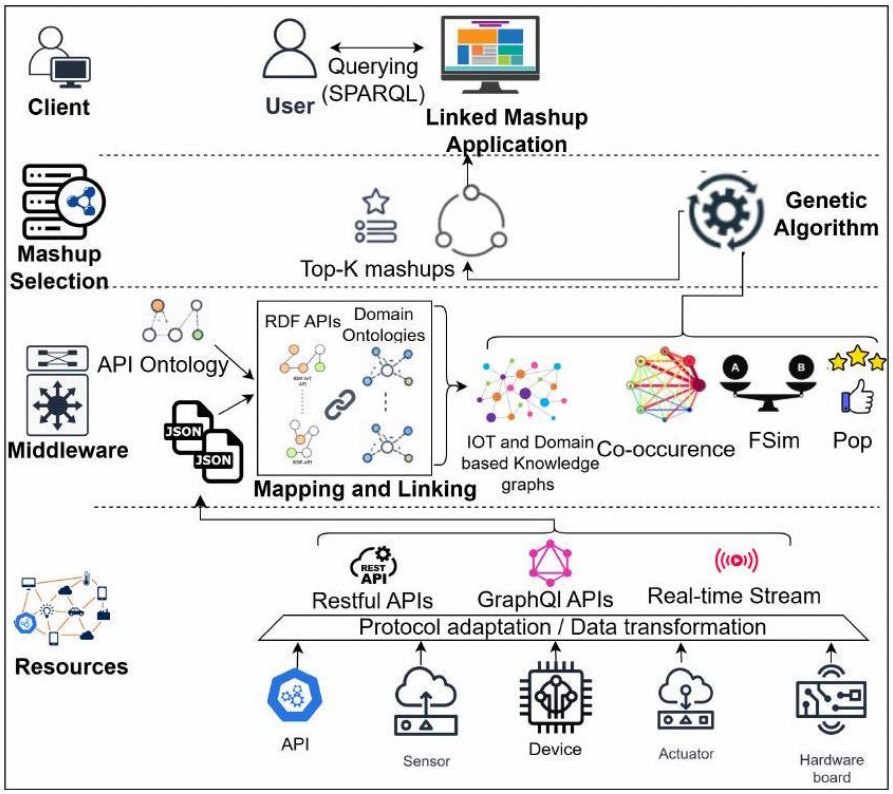


Figure 2. General architecture of the proposed mashup framework.

- Resources Layer: Includes Web APIs (e.g., mapping services), IoT devices (e.g., sensors) and other service interfaces, accessible *via* interfaces, such as GraphQL and REST. This layer supplies heterogeneous streams for mashups.

- Middleware Layer: Provides semantic interoperability by lifting raw API descriptions into RDF using our API ontology and aligning them with domain ontologies relevant to the user's request. This enhances semantic similarity and co-occurrence analysis for more accurate compositions.
- Mashup Selection Layer: Matches APIs to user requirements by evaluating similarity, co-occurrence and popularity, then applies a genetic algorithm to select the Top-K reliable compositions.
- Client Layer: Allows users to specify requirements and interact with the resulting mashups. Compositions are stored in a triple store and can be queried through a SPARQL endpoint for adaptive reuse.

Our main contributions revolve around the Middleware and Mashup Selection Layers: (1) semantic enrichment of APIs (Sub-ection 3.2) and (2) history-aware Top-K mashup composition (Sub-ection 3.3).

3.2 Lifting Web and IoT API and Mashup Descriptions into an RDF-based Model

We employed a semantic-transformation process to lift the raw data, mostly represented in JSON that describes the Web and IoT APIs and Mashups (see Figure 4 as an example) into an RDF-based representation (see Figure 5). This transformation aims at enhancing the interoperability and semantic richness of the description, making it more easily discoverable and linkable with other APIs on the Linked Data, particularly other Web and IoT APIs. Initially, we identified key components within the JSON structure, such as service name, description, functions, parameters and response properties.

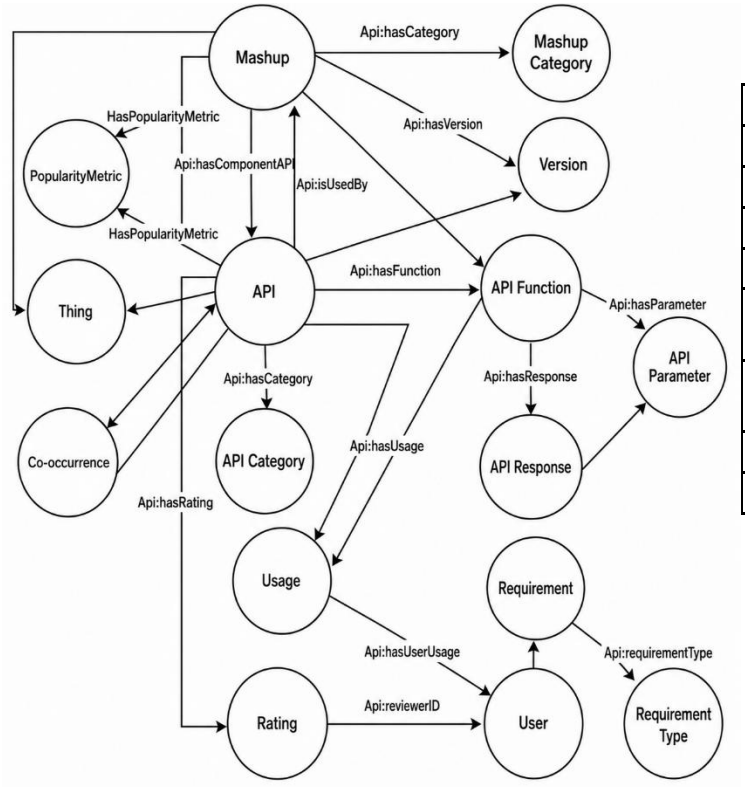


Figure 3. Proposed API-centric ontology.

Table 2. Ontology structural information measurement values.

Measurement	Value
Number of imported ontologies	2
Number of imported classes	15
Number of created classes	40
Subclass relations	17
Number of object properties	28
Number of data properties	20
Syntax language	English
Ontology engineering tool	Protégé

To do so, we developed the API-centric ontology by adopting the METHONTOLOGY ontology-development life cycle [40]. This methodology aligns with recent agile approaches that combine automated extraction with expert validation, as demonstrated in ontology development for collaborative innovation [41]. Table 2 provides measurement values of structural information of our developed ontology, while Figure 3 illustrates a portion of the ontology. The process had several stages: specification, conceptualization, formalization, implementation and maintenance. In the specification step, the goal, scope and intended users were defined, focusing on semantically representing and integrating API components and related domains for researchers and software engineers. Then, the conceptualization involved designing a conceptual model based on existing ontologies and iteratively developing the main concepts and relationships. Next, in the formalization step, the main concepts,

hierarchies and restriction axioms were represented, such as defining the API concept with its properties and relationships. Finally, the implementation is carried out using the Web Ontology Language (OWL) https://www.w3.org/OWL/ and Resource Description Framework (RDF) https://www.w3.org/RDF/, developed with Protégé and evaluated using SPARQL queries, while the maintenance ensures ongoing updates and refinements. Throughout the development of the ontology, key activities were conducted in parallel: knowledge acquisition through expert, domain-level evaluation with iterative feedback sessions and comprehensive documentation with labels and comments to describe the classes and properties. The ontology is evaluated against FAIR principles [42].

To ensure that the proposed ontology[1] is both semantically consistent and functionally adequate for supporting mashup composition, we performed a formal validation process using Protégé and its plugin-the Pellet reasoner. Consistency checking confirmed that no logical contradictions existed among class hierarchies, object properties, or data property domains and ranges, ensuring that ontology axioms were coherent. Beyond structural validation, a set of competency questions was defined to evaluate whether the ontology could effectively answer domain-relevant queries, such as: (1) Which APIs provide geolocation data usable in a travel mashup? (2) Which APIs can extract or analyze text features for content analysis? (3) Which combinations of APIs have been most frequently co-used in historical mashups? The ontology successfully resolved all competency questions using SPARQL queries over the generated RDF dataset, demonstrating its adequacy for semantic discovery and interoperability.

The connection from semantic lifting to recommendation is straightforward: user requirements are mapped to domain ontologies, which then query the RDF store for APIs the descriptions of which match these concepts. Candidate APIs are scored using three normalized metrics (range [0,1]) stored directly as RDF properties: resource:similarityValue (semantic proximity *via* LogMap), resource:coOccurrenceValue (joint usage frequency) and resource:popularityValue (aggregated user ratings). These scores feed into the reliability equation, the weight of which is optimized per request by the genetic algorithm.

For example, a user requesting a travel mashup triggers the selection of maps, weather and hotel APIs. with our API ontology in use, these APIs are lifted to RDF; domain ontologies (geolocation, weather, points of interest), then link these RDF descriptions into a unified semantic space. Multiple potential mashups-each combining one API per category-are formed. To select the best mashup, the functional similarity (computed *via* LogMap), co-occurrence and popularity (both derived from historical mashup data) of these candidates are fed into the genetic algorithm, which optimizes the six weights in the reliability equation and returns the Top-K most reliable compositions.

The resource: similarityValue property quantifies semantic proximity between services. This value is calculated using LogMap to generate ontology-based embeddings, with cosine similarity determining the normalized score in the range [0,1]. For instance, a speech-to-text API and a text-translation API might achieve a similarity value of 0.8, reflecting their complementary functions.

resource: coOccurrenceValue captures the historical frequency with which APIs appear together in successful mashups. This value is calculated as the normalized count of joint occurrences divided by the total usage frequency of both services. For example, a mapping API and a weather-service API that frequently co-appear in travel applications would receive a high co-occurrence value, indicating proven compatibility and integration feasibility.

resource: popularityValue measures service adoption and user satisfaction through aggregated ratings and usage statistics from multiple sources. This metric ranges from [0,1], where higher values indicate wider community adoption and positive user feedback. A highly-rated social-media API with extensive documentation and active developer communities would exhibit elevated popularity scores.

These three normalized values serve as critical inputs to the reliability computation in Equation 2, enabling the Genetic Algorithm to dynamically balance semantic alignment, historical compatibility and community validation during mashup optimization. By embedding these semantic-proximity measures directly into RDF descriptions, our framework enhances service discoverability and supports informed composition decisions for both developers and non-experts.

[1] Public GitHub repository: https://github.com/abderrahmane-khiat/SIMM/blob/main/Ontology/api_mashup_ontology.owl

After enriching media API descriptions using our API ontology, domain ontologies are selected based on user requirements to interlink these RDF descriptions. In the media domain, such ontologies include MPEG-7 for structural video descriptors, schema.org/VideoObject for metadata, the Audio Ontology for sound events, or FOAF for linking social-media entities. These ontologies define shared concepts and relationships that allow heterogeneous APIs-such as a video streaming API, a captioning API and a sentiment analysis API-to be linked into a coherent semantic space. For example, MPEG7:VideoSegment→hasAnnotation→ schema:Caption, or schema:VideoObject → hasFeature → schema:Sentiment. This linkage helps media-intelligent tasks, such as generating context-aware video summaries, synchronizing subtitles with visual events, or adapting content delivery based on user preferences. In this way, RDF-based enrichment combined with domain ontologies enables mashups to provide semantically robust, adaptive and personalized Intelligent Media experiences.

3.3 Top-K Reliable Mashup Composition with Popularity, Co-Occurrence and Functional Similarity

After linking different APIs, we leverage composition history through three primary functions: Co-Occ (Co-Occurrence), Pop (Popularity) and FSim (Functional Similarity). These functions are combined in our reliability formula and applied simultaneously to identify the most robust compositions to retain. We first define the three functions, followed by the overall reliability function.

```
{
  "name": "Google␣Maps␣Directions␣Service",
  "description": "Get␣driving,␣biking,␣or␣transit␣"
    "directions␣between␣locations.",
  "functions": [
    {
      "name": "getDirections",
      "parameters": [
        {
          "name": "origin",
          "type": "string",
          "description": "Starting␣location␣address␣"
            "or␣coordinates"
        },
        {
          "name": "destination",
          "type": "string",
          "description": "Ending␣location␣address␣"
            "or␣coordinates"
        },
        {
          "name": "travelMode",
          "type": "string",
          "description": "Optional:␣'driving',␣"
            "'bicycling',␣or␣'transit'"
        }
      ],
      "response": {
        "properties": {}
      }
    }
  ],
  "resources": "https://developers.google.com/maps/"
    "documentation/directions"
}
```

Figure 4. JSON description of a map API.

```
@prefix rdf:
     <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix api: <http://example.org/api/> .
@prefix ex: <http://example.org/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

ex:GoogleMapsDirectionsService a api:Service ;
    rdfs:label "Google␣Maps␣Directions␣Service" ;
    api:description "Get␣driving,␣biking,␣or␣transit␣"
        "directions␣between␣locations." ;
    api:hasFunction ex:getDirections ;
    api:resources <https://developers.google.com/maps/"
␣␣␣␣␣␣␣␣"documentation/directions> ;
    api:similarAPI [
        api:uri <https://example.com/similarAPI1> ;
        api:similarityValue "0.8"^^xsd:float
    ] ;
    api:similarAPI [
        api:uri <https://example.com/similarAPI2> ;
        api:similarityValue "0.85"^^xsd:float
    ] .

ex:getDirections a api:Function ;
    rdfs:label "getDirections" ;
    api:hasParameter ex:origin, ex:destination,
        ex:travelMode ;
    api:hasResponse ex:responseGetDirections .
```

Figure 5. RDF representation of a map API.

- Co-Occ(APIa, APIb): Defines a value that is related to the number of combinations between API a and API b. The value of Co-Occurrence is calculated by the following formula 1:

$$\text{Co} - \text{Occ}(\text{APIa}, \text{APIb}) = \frac{\text{number of compositions involving } API_a \text{ and } API_b}{\text{Total number of compositions involving } API_a \text{ or } API_b} \quad (1)$$

We did not limit ourselves to APIs, we also considered mashups that were already built. In order to distinguish between APIs and Mashups, we denoted MCo-Occ(M_a, M_b) for the Co-Occurrence calculation between two combinations of APIs.

- Pop(API_a) : Defines a value of the evaluations given by users to assess each API. To distinguish between APIs and Mashups, we denoted MPop (M_a) for the evaluations given by users to evaluate the combination of APIs. The value returned by Pop(API_a) or MPop(M_a) is a real number between 0 and 1.

- FSim(API_a, API_b) : The functional similarity used in our approach integrates the LogMap system to compute the similarity between Web and IoT APIs and compositions. LogMap [43]

is a scalable ontology-matching system with built-in reasoning and repair. It indexes input ontologies lexically and structurally, using inverted lexical indexes and external lexicons, like WordNet, to identify synonyms. It also leverages class hierarchies and employs reasoning to minimize inconsistencies in the aligned ontology. The LogMap system performs instance matching e.g., finding owl:sameAs property between linked data resources. In our research, these resources present a description of each API. LogMap shows very good results in terms of F-measure in the SPIMBENCH Benchmark series (https://oaei.ontologymatching.org/2015/im/index.html) which consists of a dataset with the purpose of determining when two RDF descriptions describe the same CreativeWork.

Similarly, we did not limit ourselves to APIs, we also considered mashups that were already built. To differentiate between APIs and Mashups, we denote $\mathrm{MFSim}(M_a, M_b)$ for the similarity computed between two combinations of APIs. We noted that the selected correspondences between APIs and compositions are added to the RDF-based API descriptor.

With multiple Web, IoT and other APIs available, applying these three functions generates four matrices (FSim, MFSim, Co-Occ, MCo-Occ) and two vectors (Pop, MPop). Together, they quantify similarity, historical co-use and user adoption, all of which are essential indicators of reliability in mashups.

To combine these indicators, we defined the Reliability Formula below. It calculates the composability between APIs and mashups, ensuring that only those compositions most likely to produce coherent and adaptive services are retained:

$$\begin{aligned}\mathrm{Reliability}(X) = {} & w_f \cdot \frac{1}{n}\sum_{i=1}^{n} \mathrm{FSim}_i + w_c \cdot \frac{1}{n}\sum_{i=1}^{n} \mathrm{CoOcc}_i + w_{fm} \cdot \frac{1}{n}\sum_{i=1}^{n} \mathrm{MFSim}_i + w_{cm} \\ & \cdot \frac{1}{n}\sum_{i=1}^{n} \mathrm{MCoOcc}_i + w_p \cdot \frac{1}{n}\sum_{i=1}^{n} \mathrm{Pop}_i + w_{pm} \cdot \frac{1}{n}\sum_{i=1}^{n} \mathrm{MPop}_i \end{aligned} \quad (2)$$

While each element denoted the APIs of a composition, $w_f, w_{fm}, w_c, w_{cm}, w_p$ and w_{pm} denoted the weights attributed respectively to the functional similarity, co-occurrence and popularity of both APIs and mashups.

The six weights are bounded between 0 and 1 and are initialized to 0.5 to give equal importance to all metrics before optimization.

The final step was to select the Top- K compositions identified by the Reliability Formula. Because the weights of FSim, Co-Occ and Pop directly determine the outcome, we employed the Genetic Algorithm (Algorithm 1) to optimize these weights. This ensures that the selected mashups are not only statistically reliable, but also contextually adaptive, producing compositions that align with user needs and semantics.

Candidate mashups are generated by, first, grouping APIs into functional categories (e.g., mapping, weather, social) based on their descriptions and the selected domain ontologies. For a user request that requires m distinct functionalities, we formed all combinations of one API from each of the m relevant categories. This yielded an initial candidate pool the size of which depends on the number of available APIs per category. From this pool, a random sub-set of $50 - 100$ candidate mashups was selected as input to the genetic algorithm to keep optimization computationally feasible. The sub-set is drawn uniformly without replacement, giving each possible combination an equal chance of being chosen. The random-number generators were not initialized with a fixed starting point. No explicit pre-filtering stage is applied; instead, the reliability scoring (Equation 2) naturally assigns low scores to low-quality combinations, so they are unlikely to appear among the Top-K recommendations. The size of the sub-set follows established guidelines for genetic-algorithm population sizes: De Jong [44] showed that $50 - 100$ individuals provide a good trade-off between diversity and convergence, a recommendation widely adopted in the GA literature. The value $K = 5$ was chosen, because it is the most commonly used setting in the API and mashup recommendation literature, offering a good balance between recall and precision. A systematic sensitivity analysis of K has not been performed.

Algorithm 1 outlines the weight-optimization process. The GA receives the randomly selected sub-set of candidate mashups, each already scored by the initial reliability function. For example, if a user

requests the creation of a travel planner, the system may combine mapping APIs (M1-M3), weather APIs (W1-W3) and hotel APIs (H1-H3) into multiple candidate mashups, such as M1-W2-H2 or M3-W3-H1. These candidates, ranked by the initial reliability function, form the pool from which the optimization sub-set is randomly sampled.

Algorithm 1. Genetic Algorithm weights optimization

Require: Subset of candidate mashups M_{sub}, Validation set V, Number of individuals N, Crossover rate C_{rate}, Mutation rate M_{rate}

Ensure: Optimized weight vector $\mathbf{w}^* = (w_f^*, w_c^*, w_p^*, w_{fm}^*, w_{cm}^*, w_{pm}^*)$

Initialize population P with N random weight vectors $\mathbf{w} = (w_f, w_c, w_p, w_{fm}, w_{cm}, w_{pm})$ where each $w_i \in [0,1]$

▷ Each individual encodes a possible weighting configuration for API- and mashup-level reliability

repeat

for each individual $\mathbf{w}$ in P **do**

Apply weights $\mathbf{w}$ to compute API-level reliability for each component API

▷ Reliability(API) $= w_f \cdot F\text{ Sim} + w_c \cdot CoOcc + w_p \cdot \text{Pop}$

Aggregate API reliabilities to compute mashup-level reliability

▷ Reliability(Mashup) $= w_{fm} \cdot F\text{ Sim}_m + w_{cm} \cdot CoOcc_m + w_{pm} \cdot \text{Pop}_m$

Rank mashups in M_{sub} by their computed reliability and select Top- K

▷ Identify the most promising mashups for validation

Evaluate selected mashups against validation set V

▷ Compute precision, recall, and F-measure relative to the validation ground truth

Compute fitness $(\mathbf{w})$ = F-measure $(\mathbf{w})$

▷ Assign the F-measure as the fitness value of the weight configuration

end for

Apply tournament selection and crossover (simulated binary) with probability C_{rate}

▷ Combine high-performing configurations to explore new solutions

Apply polynomial mutation on weight vectors with probability M_{rate}

▷ Introduce diversity by slightly altering weight values

Update population P with the new offspring

▷ Retain best-performing configurations for next generation

until F-measure improvement < 0.001 for 20 consecutive generations or maximum 100 iterations reached

▷ Stop when fitness plateaus or iteration limit is reached

return $\mathbf{w}^* = (w_f^*, w_c^*, w_p^*, w_{fm}^*, w_{cm}^*, w_{pm}^*)$

▷ Output the optimal weights for reliable and reusable mashup composition

The GA does not create new mashup compositions from scratch, but it instead refines the reliability-evaluation process by learning optimal weight configurations for the specific request at hand. At each generation, the GA:

(a) Applies the current weight configuration $\mathbf{w} = (w_f, w_c, w_p, w_{fm}, w_{cm}, w_{pm})$ (all weights in [0,1]) to compute API-level and mashup-level reliability scores for the sub-set of mashups;

(b) Performs tournament selection, simulated binary crossover (SBX) and polynomial mutation operations on the weight vectors (mutation rate $= 0.05$, crossover rate $= 0.8$);

(c) Evaluates the resulting reliability scores to compute precision, recall and F -measure as fitness metrics against a validation sub-set (20% of the candidate pool);

(d) Updates the population of weight configurations according to their fitness values until convergence criteria are met.

(e) The GA terminates when the F-measure improvement is less than 0.001 for 20 consecutive generations or a maximum of 100 generations is reached, whichever occurs first.

This iterative process establishes a dynamic balance between the semantic characteristics of APIs and their historical usage patterns: the initial candidate generation provides meaningful mashup structures based on semantic compatibility, while the GA continuously learns how to balance functional similarity, co-occurrence and popularity indicators across both API and mashup levels. The outcome is an adaptively weighted reliability model that enhances the robustness and reusability of mashup composition.

4. EXPERIMENTAL STUDY

This section details our experimental study.

4.1 Evaluation Dataset

To the best of our knowledge, there is currently no standardized benchmark for evaluating API integration and composition in service-oriented systems. Therefore, we curated an existing dataset - "Data Correction and Evolution Analysis of the ProgrammableWeb Service Ecosystem" [45] resource[2] by 1) filtering for APIs with categories relevant to common mashup domains (e.g., Mapping, Social, Weather, Media, IoT) based on keyword matching of ProgrammableWeb category tags 2) augmenting each API with a popularity metric obtained through systematic searches across multiple online API directories and developer forums (e.g., RapidAPI, API.io), providing a realistic assessment of API relevance and current usage trends. Popularity scores from different sources were normalized to the [0,1] range using Min-Max scaling; APIs with no available ratings were assigned a neutral value of 0.5. It was noted that the original dataset already includes co-occurrence of APIs. After cleaning and validation, the final curated dataset[3] contained 11,492 APIs and 7,336 mashups.

Next, to validate the performance of our system, we conducted a user test with 90 participants. Participants were asked to enter their requirements followed by a semi-structured interview. The goal of this study was to evaluate how our approach generates the best mashups that can answer different user requirements.

The dataset encompassed APIs that provide sensor data, social-media content, geospatial information, authentication and authorization services, blockchain interactions, business analytics and currency-exchange data.

4.2 Participants' Task and Evaluation Metrics

We recruited users by invitation to participate in our pilot study. The mailing list was formed by one author of this research (a student) by inviting his fellow students. In the end, we sent one round of invitations to 173 students to come up with 2 ideas of what they would like to have as a mashup system and we received 90 responses (52% response rate). Participants were aged 20-26 years, predominantly computer-science students with prior exposure to API-based systems. This sample size is adequate for an initial proof-of-concept evaluation; however, we acknowledge that the exclusive use of student participants may limit generalizability to professional developer communities. Accordingly, the findings should be regarded as a proof-of-concept rather than a definitive validation. A broader study with industry practitioners is planned for future work. It is worth mentioning that gathering free participants was very difficult and only 90 out of 173 accepted to participate. Furthermore, we did not pay participants on site or online *via* crowd platforms (e.g., MTurk) to avoid influencing the results of the evaluation.

For the task, we deployed our system on the web and we told participants that the higher their points were, the more they were promoting research. It was a form of motivation. Next, we asked them to use our system without any tutorial in order to get the first feedback about our system. The overall task (10 minutes) consisted of a 5-minute session for interacting with the system validation, followed by a 5 - minute interview. At the end, we gathered 180 systems.

Next, to validate our approach in generating mashups that best meet the user requirements, we built a ground truth by sharing the 180 systems among the participants and asking them to rate the best system (good, not specifically, not at all). The systems that received more than 10 good ratings were retained (this threshold was chosen to ensure that at least 10% of the 90 participants agreed on the quality of a system, providing a basic consensus filter). At the end, we ended up with 120 systems.

This ground truth created by participants allowed us to use measures such as precision (3), recall (4) and F-measure (5) to determine the new compositions' reliability. In the following, we define the measures Recall, Precision and F-measure used to evaluate the performance of our approach:

---

[2] Original dataset available at: https://github.com/HIT-ICES/Correted-ProgrammableWeb-dataset/tree/main/data

[3] Final dataset available at: https://github.com/abderrahmane-khiat/SIMM/tree/main/Media%20APIMashup%20dataset

- Precision: it denotes the precision level of selected resources (Web and IoT APIs), indicating the ratio of the chosen relevant APIs to the total number of chosen APIs. It is formally defined as follows:

$$\text{Precision } = \frac{R(A) \cap C(A)}{C(A)} \tag{3}$$

- Recall: it refers to the recall value of selected resources (Web and IoT APIs), indicating the ratio of the chosen relevant APIs to the total number of actual relevant APIs. It is formally defined as follows:

$$\text{Recall } = \frac{R(A) \cap C(A)}{R(A)} \tag{4}$$

- The F-measure: a tradeoff metric that balances recall and precision and is calculated using the formula:

$$F\text{-measure } = \frac{2 \times \text{ Precision} \times \text{Recall}}{\text{Precision } + \text{ Recall}} \tag{5}$$

Higher values of Recall, Precision and F-measure indicate that the system more accurately generates mashups that align with the user requirements.

All experiments were implemented in Python, running on a Windows 10 computer equipped with an Intel Core CPU at 2.8GHz, 16 GB RAM. The complete source code, including the framework implementation, baselines re-implementation and evaluation scripts, is publicly available for reproducibility under a GitHub repository[4].

4.3 Evaluation Protocol

For the evaluation of our approach, we conducted three studies:

- In the first study, we compared our approach with three systems: SSR (scores services individually, causing redundancy), CSR (groups services by function, but uses coarse cluster-level compatibility) and SPR (leverages topic/tag/history data, but assumes functional similarity ensures compatibility). Although these are API recommendation systems, their ability to compose compatible services provides a meaningful benchmark for evaluating mashup generation. To ensure a rigorous and fair comparison, we re-implemented SSR [35], CSR [36] and SPR [37] according to their original specifications. All baselines were evaluated under identical conditions using our curated dataset, with parameter settings aligned with their respective publications (e.g., Top-K values, similarity thresholds). This eliminates potential confounding factors when comparing performance metrics.

  In addition, we conducted tests of statistical significance of the comparative results using paired t-tests on the F-measure scores across each evaluation, with the null hypothesis stating no difference in performance between our approach and each baseline. We used the standard 95% confidence level and reported the resulting confidence intervals to quantify the precision and stability of the performance estimates.

- The second part of our investigation demonstrates the value of the metrics employed in our proposed solution by analyzing them both separately and in combination.
  - Functional Similarity (FSim): This baseline considers functional similarity between services based on their descriptions.
  - Popularity (POP): This method ranks services based on the number of times they have been used in mashups.
  - Co-Occurrence (CO): This baseline considers the frequency with which services appear together in mashups.

- In the third study, we conducted multiple experimental trials to evaluate how our GA algorithm performs against traditional methods in determining optimal values for the six parameters (w_f, $w_c, w_p, w_{fm}, w_{cm}, w_{pm}$), which correspond to the metrics Co-Occ, FSim, Pop, MFSim,

4 Code repository: https://github.com/abderrahmane-khiat/SIMM/tree/main/Code

MCoOcc and MPop in Equation 2. More specifically, we compared our GA algorithm with 1) static approaches that apply fixed weight values regardless of context in which all weights equal to $1/6$ and 2) empirically tuned static weights where those weights are determined through preliminary experiments, e.g., $w_f = 0.5, w_c = 0.4, w_p = 0.6, w_{fm} = 0.6, w_{cm} = 0.5, w_{pm} = 0.8$.

Our designed genetic algorithm was based on preliminary experiments and established GA literature. A population size of 100 was found to provide sufficient diversity. Similarly, the crossover rate (0.8), mutation rate (0.05) and termination criterion (20 stagnant generations) fall within commonly recommended ranges and demonstrated stable convergence across repeated runs. Additionally, we investigated the effect of varying the Top-K APIs and candidate mashups on the system's final recommendations, providing insights into how parameter tuning and selection impact the accuracy and adaptivity of mashup generation.

4.4 Results

4.4.1 Comparative Evaluation Results

We evaluated the resulting mashup compositions using our reliability formula against the ground truth. Table 3(a) presents the Precision, Recall and F-measure for our approach compared with the SSR 35, CSR [36] and SPR [37] systems. The results demonstrate that our framework achieves higher precision and recall across all metrics, indicating a significant improvement over baseline methods. Paired t-tests revealed that the improvements in F-measure were statistically significant with $p < 0.01$ for all three comparisons (*vs.* SSR: $t = 9.12, p < 0.0001$; *vs.* CSR: $t = 8.75, p < 0.0001$; *vs.* SPR: $t = 7.95, p = 0.0003$). SSR tends to produce redundant outputs by scoring services individually and CSR relies on coarse cluster-level compatibility, while SPR assumes that functional similarity alone guarantees compatibility.

Our approach systematically addresses these limitations by integrating three complementary metrics: Similarity, which captures the semantic relatability between heterogeneous APIs even when they use different specification terms; Co-Occurrence, which identifies frequently interacting APIs, reflecting compatibility, ease of integration and practical utility within a mashup; and Popularity, which reflects user-driven preferences and requirements. By combining these metrics, our system develops a more comprehensive understanding of the relationships among APIs and mashups, resulting in more accurate, adaptive and context-aware mashup recommendations. We achieved a higher recall than precision, which is common across all methods and is desirable for mashup recommendation. Our main goal was to find as many useful and compatible APIs as possible. In this context, missing a key API (false-negative) is more detrimental to mashup viability than suggesting extra, sub-optimal candidates (false-positives). Thus, our Recall-oriented performance profile aligns with the practical need for expansive discovery in adaptive ecosystems.

Table 3. Comparison results.

Metric	SPR	CSR	SSR	Our Approach
Precision	0.24	0.26	0.29	**0.37**
Recall	0.44	0.49	0.56	**0.68**
F-measure	0.31	0.34	0.38	**0.48**

(a) Performance comparison with baseline methods. Improvements are statistically significant ($p < 0.01$). Our approach achieves 26% improvement in F -measure.

FSim	Pop	Co-Occ	Combined
0.30	0.29	0.33	**0.37**
0.55	0.52	0.56	**0.68**
0.39	0.37	0.42	**0.48**

(b) Individual metric contributions *versus* combined performance

4.4.2 Parameter Combination Study

Table 3(b) illustrates the performance improvements achieved by combining the Similarity, Co-Occurrence and Popularity metrics compared to using each metric individually. These results show that each metric captures a distinct aspect of mashup quality. By simultaneously considering fine-grained compatibility, semantic relatability and user-driven relevance, our approach achieved superior performance.

We further investigated the effect of parameter application order on mashup selection, considering two strategies: sequential and parallel.

- Sequential approach: Parameters were applied in the following order Co-Occurrence → Similarity → Popularity, meaning that the selection of mashups was first influenced by API interaction patterns, then by semantic similarity and finally by user-oriented popularity.
- Parallel approach: All three metrics were applied simultaneously, weighted equally in the reliability formula and evaluated using Precision, Recall and F-measure.

The results indicate that the parallel approach provides more consistent and balanced outcomes, with precision and recall values generally 10% - 15% higher than the sequential approach. By treating all metrics simultaneously, the system avoids bias toward any single factor and produces more adaptive and accurate mashup recommendations.

4.4.3 Adaptive *vs.* Static Weight Optimization

This evaluation demonstrates the advantage of our genetic algorithm-based adaptive weighting approach compared to traditional static weighting methods. Unlike static approaches that apply fixed weight values regardless of context, our adaptive method dynamically computes optimal weight configurations for each individual mashup request through evolutionary optimization.

The key distinction lies in the optimization process: while static approaches use predetermined weights (typically equal weighting or values empirically tuned through preliminary experiments), our dynamic approach employs a genetic algorithm that, for every mashup request, explores the weight space, evaluating configurations against a validation sub-set to identify optimal balances between functional similarity, co-occurrence and popularity metrics specific to that request.

Table 4 presents the performance comparison between different weighting strategies that we used in our approach. The genetic-algorithm optimization consistently outperformed all static approaches, achieving superior F -measure (0.48) through balanced improvements in both precision (0.37) and recall (0.68).

Table 4. Performance comparison: Static weighting *vs.* Per-request genetic algorithm optimization.

Weighting Strategy applied on Our Approach	Precision	Recall	F-measure
Equal weighting (all weights = $1/6$)	0.31	0.59	0.41
Empirically tuned static weights	0.33	0.60	0.43
Genetic algorithm optimization	$\mathbf{0.37}$	$\mathbf{0.68}$	$\mathbf{0.48}$

The comparative evaluation revealed two findings:

- The optimal weight configuration varied significantly across different mashup requests, even within the same application domain. For instance, requests for travel planning mashups showed weights of $w_f \approx 0.85, w_c \approx 0.45, w_p \approx 0.90$, while social-media aggregation requests converged to $w_f \approx 0.75, w_c \approx 0.60, w_p \approx 0.95$. This is because the importance of each similarity, popularity or co-occurrence is different for each request. This fine-grained adaptation is only possible using genetic-algorithm optimization.
- Static approaches failed to capture these request-specific variations, leading to sub-optimal recommendations. Our adaptive method demonstrated over 10% better performance for novel or complex requests where static approaches make little use of historical patterns.

The results show that the optimal configurations of our genetic algorithm were highly request-dependent

with weights adjusting based on:

- The specificity and novelty of user requirements;
- The availability and characteristics of relevant APIs;
- The density of historical usage patterns in the target domain;
- The balance between semantic compatibility and practical integration concerns.

Regarding computational scalability, the GA optimization completed within 2-3 seconds per request on our test hardware. For larger API ecosystems, candidate generation can be accelerated using indexing structures to pre-filter APIs before semantic matching, ensuring that the approach remains practical for real-time applications.

5. CONCLUSION

In this research paper, we presented a novel history-aware semantic framework for generating reliable and reusable mashups that addresses key challenges in heterogeneous API integration and composition. To achieve this, we first developed an API ontology and JSON-to-RDF mapping for semantic enrichment of APIs to harmonize APIs and compositions for better API discovery. Second, we incorporated three complementary metrics-functional similarity, co-occurrence and popularity-within RDF-based descriptions in order to learn the historical composition patterns. Finally, we employed a genetic algorithm that performs dynamic weight optimization to produce top- k reliable mashup recommendations.

Experimental evaluation on a curated dataset of 11,492 APIs and 7,336 mashups demonstrated statistically significant improvements of 26% in F-measure over state-of-the-art baselines (SSR, CSR, SPR), with all comparisons achieving $p < 0.01$.

The results confirm that by unifying semantic interoperability with adaptive, context-aware reliability analysis, our approach promotes the long-term robustness and reusability of mashup compositions. We emphasize that our contribution to sustainability is limited to enhancing reusability-a critical, but narrow, dimension of the broader sustainability concept. The use of Intelligent Media mashups in our experiments serves as a representative case study; the framework itself is domain-agnostic and can be applied to general IoT and Web API composition scenarios. Generalizability to real-world API ecosystems remains to be established. Furthermore, in future work, we plan to address some limitations, such as conducting a systematic sensitivity analysis of K and initializing the random-number generators with a fixed starting point to ensure exact reproducibility, in addition to exploring more promising directions through machine-learning techniques enhanced by ontological knowledge leveraging structured semantics to further improve the accuracy, reliability and reusability of service compositions across application domains.

REFERENCES

[1] R. E. Roth, R. G. Donohue, C. M. Sack, T. R. Wallace and T. M. Buckingham, "A Process for Keeping Pace with Evolving Web Mapping Technologies," Cartographic Perspectives, no. 78, pp. 2552, 2014.

[2] Y. Qin et al., "When Things Matter: A Survey on Data-centric Internet of Things," Journal of Network and Computer Applications, vol. 64, pp. 137-153, 2016.

[3] E. Borgia, "The Internet of Things Vision: Key Features, Applications and Open Issues," Computer Communications, vol. 54, pp. 1-31, 2014.

[4] Y. Maaravi and B. Heller, "Digital Innovation in Times of Crisis: How Mashups Improve Quality of Education," Sustainability, vol. 13, no. 13, p. 7082, 2021.

[5] B. Beemer and D. Gregg, "Mashups: A Literature Review and Classification Framework," Future Internet, vol. 1, no. 1, pp. 59-87, 2009.

[6] F. Daniel and M. Matera, Mashups-Concepts, Models and Architectures, E-ISSN: 2197-974X, Springer Berlin, Heidelberg, 2014.

[7] A. K. Kalou and D. A. Koutsomitropoulos, "Towards Semantic Mashups: Tools, Methodologies and State of the Art," International Journal of Information Retrieval Research, vol. 5, no. 2, pp. 1-25, 2015.

[8] W. Pan, X. Xu, H. Ming and C. K. Chang, "Clustering Mashups by Integrating Structural and Semantic Similarities Using Fuzzy ahp," Int. J. of Web Services Research, vol. 18, no. 1, pp. 34-57, 2021.

[9] S. Handschuh, "Semantic Annotation of Resources in the Semantic Web," Chapter in Book: Semantic Web Services: Concepts, Technologies and Applications, Semantic Annotation of Resources in the Semantic Web, pp. 135-155, DOI: 10.1007/3-540-70894-4_5, 2007.

[10] P. Rodriguez-Mier et al., "An Integrated Semantic Web Service Discovery and Composition Framework," IEEE Trans. on Services Computing, vol. 9, no. 4, pp. 537-550, 2015.
[11] D. Bonetta, A. Peternier, C. Pautasso and W. Binder, "S: A Scripting Language for High-Performance Restful Web Services," ACM Sigplan Notices, vol. 47, no. 8, pp. 97-106, 2012.
[12] R. Alarcon, E. Wilde and J. Bellido, "Hypermedia-driven Restful Service Composition," Proc. of ICSOC 2010 Workshops, LNCS 6568, pp. 111–120, Springer, 2011.
[13] D. Cooper et al., Internet Engineering Task Force, Fremont, USA, 2008.
[14] R. Verborgh et al., "Description and Interaction of Restful Services for Automatic Discovery and Execution," Proc. of the Int. Workshop on Advanced Future Multimedia Services (AFMS), pp. 1-10, [Online], available: https://backoffice.biblio.ugent.be/download/2003291/2003308, 2011.
[15] D. John and M. Rajasree, "Restdoc: Describe, Discover and Compose Restful Semantic Web Services Using Annotated Documentations," Int. J. of Web & Semantic Technology, vol. 4, no. 1, p. 37, 2013.
[16] J. Koch, C. Velasco and P. Ackermann, "HTTP Vocabulary in RDF 1.0," W3C Working Draft, World Wide Web Consortium, 2011.
[17] R. Milner, "The Polyadic π-calculus: A Tutorial," Proc. of Logic and Algebra of Specification, in Book: NATO ASI Series, vol. 94, pp. 203-246, Springer, 1993.
[18] A. G. Hernández and M. N. M. García, "A Formal Definition of Restful Semantic Web Services," Proc. of the 1st Int. Workshop on Restful Design, pp. 39-45, DOI: 10.1145/1798354.179838, 2010.
[19] M. Dojchinovski and T. Vitvar, "Linked Web APIs Dataset," Semantic Web, vol. 9, no. 4, pp. 381-391, 2018.
[20] S. Omelkova and P. Küngas, "A Linked Data Model for Web API-s," Proc. of the 14th Int. Conf. in Business Informatics Research, pp. 48-63, DOI: 10.1007/978-3-319-21915-8_4, Springer, Estonia, 2015.
[21] F. Ardjani, D. Bouchiha and M. Malki, "An Approach for Discovering and Maintaining Links in RDF Linked Data," Int. Journal of Modern Education and Computer Science, vol. 9, no. 3, p. 56, 2017.
[22] J. Lathem et al., "SA-REST and (S)mashups: Adding Semantics to RESTful Services," Proc. of the IEEE Int. Conf. on Semantic Computing (ICSC 2007), pp. 469-476, Irvine, CA, USA, 2007.
[23] A. Malki, S. M. Benslimane, "Building Semantic Mashup," Proc. of ICWIT, pp. 40-49, [Online], Available: https://ceur-ws.org/Vol-867/Paper5.pdf, 2012.
[24] A. H. Ngu, M. P. Carlson, Q. Z. Sheng and H.-y. Paik, "Semantic-based Mashup of Composite Applications," IEEE Trans. on Services Computing, vol. 3, no. 1, pp. 2-15, 2010.
[25] N. Heath, "How IBM's Node-RED is Hacking Together the Internet of Things," TechRepublic, vol. 14, 2014.
[26] E. Prudhommeaux, "SPARQL Query Language for RDF," W3C Recommendation, [Online], Available: http://www.w3.org/TR/rdf-sparql-query/, 2008.
[27] A. Khattab et al., "Enhancing User Experience in IoT Mashup Using Semantic Technology," Proc. of the 2018 IEEE Global Conf. on Internet of Things (GCIoT), pp. 1-6, Alexandria, Egypt, 2018.
[28] S. Kim and I.-Y. Ko, "A Conversational Approach for Modifying Service Mashups in IoT Environments," Proc. of the 2022 CHI Conf. on Human Factors in Computing Systems (CHI '22), Article no. 345, pp. 1-16, DOI: 10.1145/3491102.351765, 2022.
[29] A. M. R. AlSobeh, S. AlShattnawi, A. Jarrah and M. M. Hammad, "WeaveSim: A Scalable and Reusable Cloud Simulation Framework Leveraging Aspect-oriented Programming," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 6, no. 2, pp. 182-201, 2020.
[30] M. Boulakbech, N. Messai, Y. Sam and T. Devogele, "Configuration Approach for Personalized Travel Mashup," Concurrency and Computation: Practice and Experience, vol. 35, no. 11, 2023.
[31] Y. Xu, Z. Wang, H. Gao, Z. Jiang, Y. Yin and R. Li, "Towards Machine-learning-driven Effective Mashup Recommendations from Big Data in Mobile Networks and the Internet-of-Things," Digital Communications and Networks, vol. 9, no. 1, pp. 138-145, 2023.
[32] D. Cardillo et al., "The Art of Video Mashup: Supporting Creative Users with an Innovative and Smart Application," Multimedia Tools and Applications, vol. 53, no. 1, pp. 1-23, 2011.
[33] A. Troudi, C. A. Zayani, S. Jamoussi and I. Amous, "A New Social Media Mashup Approach," Proc. of Intelligent Systems Design and Applications Conf. (ISDA2016), Part of the Book Series: Advances in Intelligent Systems and Computing, vol. 557, pp. 677-686, 2017.
[34] L. Zhang et al., "JMF-SS: Joint Matrix Factorization for Web API Recommendation with Mashup-Mashup Similarities and API-API Similarities," Proc. of the 2024 27th IEEE Int. Conf. on Computer Supported Cooperative Work in Design (CSCWD), pp. 1219-1224, Tianjin, China, 2024.
[35] W. Gao and J. Wu, "A Novel Framework for Service Set Recommendation in Mashup Creation," Proc. of the 2017 IEEE Int. Conf. on Web Services (ICWS), pp. 65-72, Honolulu, HI, USA, 2017.
[36] B. Xia et al., "Category-aware API Clustering and Distributed Recommendation for Automatic Mashup Creation," IEEE Trans. on Services Computing, vol. 8, no. 5, pp. 674-687, 2014.
[37] J. Cao, Y. Lu and N. Zhu, "Service Package Recommendation for Mashup Development Based on a Multi-level Relational Network," Proc. of the 14th Int. Conf. on Service-Oriented Computing (ICSOC 2016), pp. 666-674, Banff, AB, Canada, October 10-13, 2016.

[38] A. Khiat, M. Benaissa and M. A. Belfedhal, "STRIM Results for OAEI 2015 Instance Matching Evaluation," Proc. of the 10th Int. Workshop on Ontology Matching OM-15 Collocated with the 14th Int. Semantic Web Conference (ISWC-2015), pp. 208-215, Bethlehem, USA, 2015.

[39] S. J. R. Méndez et al., "J2RM: An Ontology-based JSON-to-RDF Mapping Tool," ISWC (Demos/Industry), pp. 368-373, [Online], Available: https://ceur-ws.org/Vol-2721/paper593.pdf, 2020.

[40] M. Fernández-López, A. Gómez-Pérez and N. Juristo, "Methontology: From Ontological Art towards Ontological Engineering," Proc. of the Ontological Engineering (AAAI-97), [Online], Available: https://cdn.aaai.org/Symposia/Spring/1997/SS-97-06/SS97-06-005.pdf, USA, 1997.

[41] F. Kharbat, A. Alshawabkeh and M. Sharairi, "A Research-based Ontology for Collaborative Innovation: A Methodology Leveraging AI and Domain Expert Knowledge," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 10, no. 03, pp. 265-280, 2024.

[42] D. Garijo and M. Poveda-Villalón, "Best Practices for Implementing FAIR Vocabularies and Ontologies on the Web," Chapter in Book: Applications and Practices in Ontology Design, Extraction and Reasoning, pp. 39-54, DOI: 10.3233/SSW200034, IOS Press, 2020.

[43] E. Jiménez-Ruiz and B. Cuenca Grau, "LogMap: Logic-based and Scalable Ontology Matching," Porc. of the10th Int. Semantic Web Conf. (ISWC2011), Part of the book series: Lecture Notes in Computer Science, vol. 7031, pp. 273-288, Germany, 2011.

[44] K. A. De Jong, An Analysis of the Behavior of a Class of Genetic Adaptive Systems, Thesis, University of Michigan, USA, 1975.

[45] M. Liu, Z. Tu, Y. Zhu, X. Xu, Z. Wang and Q. Z. Sheng, "Data Correction and Evolution Analysis of the Programmable Web Service Ecosystem," Journal of Systems and Software, vol. 182, p. 111066, 2021.

ملخص البحث:

يشكّل العدد المتزايد من واجهات برمجة تطبيقات الويب وإنترنت الأشياء غير المتجانسة تحدّياً جوهرياً لموثوقية التّطبيقات الحديثة وإمكانية إعادة استخدامها، حيث تعاني أُطُر تركيب واجهات برمجة التّطبيقات الحالية من التّجزئة الدّلالية وإهمال أنماط الاستخدامات السابقة. ولمعالجة هذه المشكلة، نقدّم إطاراً دلالياً مراعياً للتّاريخ، يعزّز متانة تركيبات الخدمات على المدى الطّويل، من خلال دمْج واجهات برمجة التّطبيقات في مخطط بياني معرفي موحّد عبر أنطولوجيا مخصّصة تنسّق البيانات والخدمات غير المتجانسة، من خلال التّجريد الدّلالي ومواءمة مستوى المجال. كذلك يتمّ تعزيز موثوقية التّركيب من خلال دمج بيانات التّركيب التّاريخية الّتي تسجّل التّواجد المشترك والشّعبية. حيث تعمل خوارزمية جينية على تحسين أوزان هذه المقاييس ديناميكياً لاختيار أفضل تركيبات الدّمج الأكثر موثوقية.

ويُظهر تقييم شامل على معيارٍ واسع النّطاق (11492 واجهة برمجة تطبيقات و 7336 عملية دمْج) تحسُّناً ذا دلالة إحصائية بنسبة 26% في موثوقية التّطبيقات المركبة مقارنة بالنماذج المرجعية. ومع أنّ هذا العمل يركّز على عمليات دمْج الوسائط الذّكية كمجال تطبيقٍ نموذجي، فإنّ الإطار المقترح قابل للتّطبيق على تركيب واجهات برمجة تطبيقات الويب وإنترنت الأشياء بشكلٍ عامّ. ويوفّر دمج التّجريد الدّلالي وتحليل التّركيبات السابقة والتّحسين حلّاً فعّالاً لإنشاء تطبيقات خدماتٍ مرنة وقابلة للتّكيُّف وذات موثوقية عالية، ممّا يقدّم منهجاً مفيداً لدعم قابلية التّكيُّف وإعادة استخدام أنظمة الخدمات المعقّدة على المدى الطّويل.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

ANALYTICAL INSIGHTS INTO THE AVERAGE BLOCK ERROR RATE OF FLUID ANTENNA SYSTEMS UNDER SHORT-PACKET CONSTRAINTS

Chi-Bao Le and Tin Phu Tran

(Received: 22-Apr.-2026, Revised: 12-Jun.-2026, Accepted: 29-Jun.-2026)

ABSTRACT

Fluid antenna systems (FASs) offer immense spatial diversity within compact spaces. However, existing performance analyses predominantly rely on Shannon's capacity bounds, assuming infinite blocklength transmissions. To address the strict requirements of emerging Internet-of-Things (IoT) and ultra-reliable low-latency communications (URLLCs), this paper investigates a point-to-point FAS under short-packet communication (SPC) constraints. We rigorously derive closed-form analytical expressions for the average block error rate (BLER), effective system throughput and reliability. Furthermore, we conduct an asymptotic analysis in the high signal-to-noise ratio (SNR) regime, mathematically revealing that the system's diversity order is strictly determined by the rank of the spatial correlation matrix. Extensive Monte Carlo simulations validate the high accuracy of our proposed analytical models, demonstrating that increasing the number of fluid ports drastically reduces the BLER.

KEYWORDS

Fluid antenna systems, Short-packet communications, Block error rate, Ultra-reliable low-latency communications, Throughput, Reliability.

1. INTRODUCTION

The rapid proliferation of the Internet-of-Things (IoT) and the stringent demands of ultra-reliable and low-latency communications (URLLC) have driven a paradigm shift in wireless network design [1][2][3]. In emerging applications, such as industrial automation, remote healthcare and autonomous systems, devices typically exchange short control commands or status updates rather than large data payloads. These scenarios require transmitting small data packets within strict millisecond-level latency constraints while ensuring near-perfect reliability [4].

Under such strict latency requirements, the classic Shannon capacity theory, which assumes infinitely long data blocks to achieve an arbitrarily low error rate, becomes inadequate [5]. For short-packet communications (SPCs), the blocklength is finite, making the probability of decoding errors non-negligible. This fundamental physical limitation is addressed by the finite blocklength information theory pioneered by [6]-[8], which mathematically characterizes the block error rate (BLER) for short packets. Consequently, the research community has extensively re-evaluated various wireless technologies, ranging from massive multiple-input multiple-output (MIMO) [9]-[11] to intelligent reflecting surfaces [12]-[14], strictly through the lens of SPC. Furthermore, while initial studies evaluated short-packet communications assuming independent fading channels, recent advancements have extended the finite blocklength-information theory to correlated fading environments [15] to better contextualize practical deployment scenarios.

Parallel to these theoretical advancements, the fluid antenna system (FAS) has emerged as a revolutionary hardware technology for compact devices [16]-[18]. Traditional MIMO systems require physical separation (typically at least half a wavelength) between antennas to ensure independent fading channels, which is impractical for small form-factor devices due to severe mutual coupling [19]. FAS overcomes this physical limitation by utilizing liquid metal or reconfigurable pixel structures, allowing a single radiating element to dynamically switch its position to the most favorable location (port) within a highly confined space [20]-[21]. By exploiting the spatial variations of the fading channel, FAS achieves significant spatial diversity without requiring multiple bulky and power-hungry radio frequency (RF) chains.

C.-B. Le (Corresponding Author) and T. P. Tran are with Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam. Emails: lechibao@tdtu.edu.vn and phutrantin@tdtu.edu.vn

While a growing body of literature has investigated the theoretical performance limits of FAS, typically employing Jakes' fading model to characterize the strong spatial correlation among densely packed ports [22], these studies predominantly rely on the infinite blocklength assumption. Very recently, the critical interplay between FAS and SPC has begun to attract attention. For instance, the authors in [23] investigated a UAV-enabled relaying network equipped with FAS under short-packet constraints. Despite this pioneering effort, their work primarily targets energy-efficiency optimization within a complex relaying topology under Nakagami-m fading. Beyond this, other recent studies, such as [24], have investigated the performance of FAS under finite blocklength constraints in multi-user setups. Currently, a comprehensive, exact and tractable analytical framework characterizing the reliability and throughput of a baseline point-to-point (P2P) FAS link under SPC constraints is still missing from the literature, even though such a baseline setup serves as the essential building block for more complex networks.

Motivated by this critical research gap, this paper comprehensively analyzes the delivery performance of a P2P FAS link operating under strict SPC constraints. The main contributions of our work are summarized as follows:

- We formulate a comprehensive mathematical system model that elegantly integrates the highly correlated spatial fading environment of the FAS with the finite blocklength constraints of SPC.
- We derive highly accurate, closed-form expressions for the exact average BLER, the effective system throughput and the overall system reliability. This is achieved by exploiting the inclusion-exclusion principle and eigenvalue decomposition (EVD) to evaluate the cumulative distribution function (CDF) of the effective SNR, combined with a linear approximation to bypass the analytically complex Gaussian Q-function.
- To gain deeper fundamental insights, we conduct an asymptotic analysis of the system performance in the high-SNR regime. This rigorous approach explicitly reveals the exact diversity order of the FAS link, mathematically demonstrating how the rank of the spatial correlation matrix fundamentally dictates the achievable reliability.
- We investigate the optimal design of blocklength allocation to maximize the effective system throughput under strict reliability and latency boundaries, providing vital operational insights for the physical implementation of short-packet FAS.
- We provide extensive Monte Carlo simulations to definitively validate the accuracy of our derived analytical models. The numerical results clearly illustrate the fundamental trade-offs among the number of fluid ports, the physical antenna space and the short-packet parameters in determining the overall system performance.

The rest of this paper is organized as follows. Section 2 describes the physical transmission framework and the spatially correlated channel model of the FAS. Section 3 presents the core mathematical derivations for the exact and asymptotic average BLER, as well as the effective system throughput and reliability, followed by an optimized short-packet blocklength design framework. Section 4 provides comprehensive numerical results to validate our proposed analytical models and finally, Section 5 concludes the paper.

Notation: $\Pr(\cdot)$ denotes probability; $\mathbb{E}\{\cdot\}$ is the expectation operator; $J_0(\cdot)$ is the zero-order Bessel function of the first kind; $Q(\cdot)$ is the standard Gaussian Q-function; $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and conjugate transpose operations, respectively.

2. System Model

2.1 Signal Transmission and FAS Topology

The fundamental limits of fluid antenna systems (FASs) have been actively investigated in recent literature. For instance, a rigorous geometric framework connecting the spatial diversity order directly to the rank of the correlation matrix was established in [25]. Furthermore, the practical deployment of FASs in 6G ultra-reliable communications must account for practical hardware non-idealities, such as the inherent trade-off between the total port count and the effective transmission blocklength caused by non-negligible port-switching delays [26]. More recently, the performance limits of FASs have been

extended to cooperative relay topologies under finite-blocklength constraints, including unmanned aerial vehicle (UAV) networks [27].

Consider a point-to-point wireless link comprising a single-antenna transmitter and a receiver equipped with a one-dimensional fluid antenna structure. The physical space allocated for the fluid antenna has a normalized length of L_s (measured in terms of the carrier wavelength). For digital-processing tractability, we assume that the receiver uniformly discretizes this continuous space into K predefined ports along L_s. Due to hardware constraints, the receiver possesses only a single active RF chain. However, it is capable of dynamically switching this RF chain to whichever of the K ports currently offers the highest signal quality.

Suppose that the transmitter sends a signal x with an average transmit power of P. The received baseband signal at the k-th port ($k \in \{1,2,\dots,K\}$) is given by:

$$y_k = \sqrt{P} h_k x + n_k, \tag{1}$$

where $h_k \sim \mathcal{CN}(0,1)$ is the complex Rayleigh flat-fading channel coefficient between the transmitter and the k-th port and $n_k \sim \mathcal{CN}(0,\sigma^2)$ represents the additive white Gaussian noise (AWGN) at the receiver with variance σ^2.

The instantaneous SNR at the k-th port is defined as $\gamma_k = \rho|h_k|^2$, where $\rho = P/\sigma^2$ is the average transmit SNR. Since the FAS controller aims to maximize the system performance, the effective SNR available for decoding the incoming packet is the maximum across all available ports:

$$\Gamma = \max_{k\in\{1,\dots,K\}} \gamma_k. \tag{2}$$

It is worth noting that the formulation in (2) relies on the assumptions of perfect channel-state information (CSI) and instantaneous port selection. In practical FAS deployments, however, hardware non-idealities, such as channel estimation overhead and mechanical/electronic RF switching delays, present notable constraints [26]. Crucially, since the system operates under SPC with a strictly limited finite blocklength, any non-negligible switching delay directly penalizes the system by reducing the effective number of channel uses available for data payload transmission. While a comprehensive analysis of these hardware impairments is beyond the scope of this baseline study, the idealized model under perfect CSI serves as an indispensable theoretical benchmark that establishes the fundamental performance upper bounds, which is a widely adopted approach in recent short-packet multi-antenna literature [34].

2.2 Spatial Correlation Characterization

Since the K ports are densely packed into a confined physical space L_s, the channel fading coefficients $h_1, h_2, \dots, h_K$ are highly correlated. Assuming a rich scattering environment, which is typical in dense IoT deployments, the spatial correlation between the fading envelopes of the i-th and j-th ports can be characterized by Jakes' fading model [28]. The spatial correlation coefficient is mathematically expressed as:

$$r_{i,j} = J_0\left(2\pi \frac{|i-j|}{K-1} L_s\right). \tag{3}$$

These pairwise coefficients constitute a $K \times K$ spatial correlation matrix, denoted as $\mathbf{R}$, where its (i,j)-th entry is $[\mathbf{R}]_{i,j} = r_{i,j}$. We explicitly acknowledge that Jakes' model relies on an idealized isotropic scattering environment with a uniform angle of arrival (AoA), which constitutes a practical limitation, as it may not fully capture directional non-isotropic scattering, shadow fading or strong line-of-sight (LoS) components common in dense, cluttered IoT deployments. Nevertheless, the adoption of Jakes' model is rigorously justified, because it serves as the foundational mathematical benchmark in fluid-antenna literature [22], [29], enabling a tractable analytical baseline. Crucially, from a physical standpoint, isotropic scattering yields the highest possible spatial correlation among closely spaced ports. Therefore, analyzing the system under Jakes' model provides a conservative, robust "worst-case" performance baseline; if the FAS link can meet the strict reliability constraints under this extreme correlation, it will perform even better in realistic environments with directional scattering or LoS paths where correlation is naturally lower.

Directly analyzing the maximum of highly correlated random variables in (2) is mathematically

intractable. To facilitate the statistical analysis, we decouple these correlated channels into a set of independent ones by applying the eigenvalue decomposition (EVD) to the correlation matrix [30]-[31].

$$\mathbf{R} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^H, \tag{4}$$

where $\mathbf{U}$ is a unitary matrix containing the eigenvectors and $\boldsymbol{\Lambda} = \mathrm{diag}(\mu_1, \mu_2, \dots, \mu_K)$ is a diagonal matrix of the corresponding eigenvalues. The eigenvalues are arranged in a descending order, i.e., $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_K \geq 0$. This mathematical transformation decouples the highly correlated fluid antenna channel vector $\mathbf{h}$ into a set of independent virtual variables *via* the linear relation $\mathbf{h} = \mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{g}$, where $\mathbf{g} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$ represents a vector of independent and identically distributed (i.i.d.) standard complex Gaussian random variables. Because $\mathbf{U}$ is a unitary matrix (i.e., $\mathbf{U}^H\mathbf{U} = \mathbf{I}_K$), this transformation completely diagonalizes the spatial correlation matrix while strictly preserving the total statistical properties and physical energy of the channel [28], [30]. Following this rigorous justification, retaining only the S dominant eigenvalues allows us to systematically reduce the system's dimensionality to its effective rank, ensuring mathematical tractability while retaining the underlying physical spatial-correlation characteristics [29].

Physically, when the space L_s is sufficiently small and the number of ports K is large, the matrix $\mathbf{R}$ becomes severely rank-deficient. Consequently, many of the smaller eigenvalues approach zero due to the extreme spatial correlation [29]. Let $S \leq K$ denote the effective rank of $\mathbf{R}$, meaning that there are only S dominant, non-zero eigenvalues $(\mu_1, \dots, \mu_S)$. This rank S dictates the true spatial degrees of freedom in the system and plays a pivotal role in the subsequent asymptotic analysis.

To formally quantify the effective rank S, we adopt a strict cumulative energy-fraction criterion. Specifically, the effective rank S is mathematically defined as the minimum number of dominant eigenvalues required to capture a significant fraction τ of the total spatial channel energy, expressed as:

$$S = \arg\min_{s} \left\{ \frac{\sum_{m=1}^{s} \mu_m}{\sum_{k=1}^{K} \mu_k} \geq \tau \right\}, \tag{5}$$

where τ is a pre-defined threshold close to unity (e.g., $\tau = 99.9\%$). This criterion guarantees that the truncated smaller eigenvalues have a mathematically negligible impact on the overall distribution accuracy while maintaining a rigorous framework for the asymptotic diversity-order analysis.

3. Performance Analysis: Average BLER

In this section, we derive the statistical properties of the maximum SNR Γ. Subsequently, we incorporate the finite blocklength Q-function to obtain the exact average BLER and its asymptotic bound.

3.1 Exact Statistical Characterization of Γ

By leveraging the EVD transformation in (C.5) and the inclusion-exclusion principle, the CDF of the effective SNR Γ for the P2P FAS can be tightly approximated as a finite sum of exponential functions [29]:

$$F_\Gamma(\gamma) \approx 1 + \sum_{k=1}^{S} (-1)^k \sum_{\mathcal{N}_k} \exp\left(-\frac{\gamma}{\rho}\Delta_k\right), \tag{6}$$

where S is the effective rank of $\mathbf{R}$ and $\sum_{\mathcal{N}_k}$ denotes the summation over all possible combinations of k distinct indices $(n_1, n_2, \dots, n_k)$ chosen from the set $\{1,2, \dots, S\}$. The term Δ_k represents the sum of the reciprocals of the chosen eigenvalues:

$$\Delta_k = \sum_{j=1}^{k} \frac{1}{\mu_{n_j}}. \tag{7}$$

3.2 Average BLER Formulation

We now incorporate the short-packet constraints. Suppose that the transmitter sends a packet containing D information bits over a finite blocklength of L channel uses, yielding a nominal coding rate of $R_c = D/L$ (bits per channel use).

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

According to the finite-blocklength information theory [32], the instantaneous BLER for a specific SNR realization Γ is closely approximated by the Gaussian Q-function

$$\epsilon(\Gamma) \approx Q\left(\sqrt{\frac{L}{V(\Gamma)}}(\log_2(1+\Gamma) - R_c)\right), \tag{8}$$

where $V(\Gamma) = (1-(1+\Gamma)^{-2})(\log_2 e)^2$ is the channel dispersion [33]. To evaluate the system reliability, the average BLER is obtained by taking the mathematical expectation of (8) over the distribution of Γ:

$$\bar{\epsilon} = \mathbb{E}[\epsilon(\Gamma)] = \int_0^\infty \epsilon(x) f_\Gamma(x) dx \tag{9}$$

where $f_\Gamma(x)$ is the probability-density function (PDF) of Γ.

Evaluating (9) directly is mathematically intractable due to the complex Q-function. To resolve this, we adopt a widely accepted linear approximation for the Q-function [23]. Let $\theta = 2^{R_c} - 1$ be the target SNR threshold required to support the rate R_c. The Q-function can be linearized around θ as:

$$\epsilon(x) \approx \begin{cases} 1, & x < a \\ \frac{1}{2} - \beta(x-\theta), & a \le x \le b \\ 0, & x > b \end{cases} \tag{10}$$

where the slope β and the boundaries a, b are defined as:

$$\beta = \sqrt{\frac{L}{2\pi V(\theta)}}, \tag{11}$$

$$a = \theta - \frac{1}{2\beta}, b = \theta + \frac{1}{2\beta}. \tag{12}$$

To justify the accuracy of this piece-wise linear model and quantify the approximation error, we refer to the analytical-integration framework recently established in [34]. Specifically, by applying integration by parts to the average BLER expectation, the piece-wise boundaries logically cancel out the CDF evaluations at the limits. Similar to the exact formulation derived in [34], Eq. (11), the average BLER elegantly simplifies to:

$$\bar{\epsilon} = \int_0^\infty \epsilon(x) f_\Gamma(x) dx \approx \beta \int_a^b F_\Gamma(x) dx \tag{13}$$

The maximum approximation error occurs only near the transition boundaries a and b. In the targeted ultra-reliable regime characterized by high average SNR ($\rho \to \infty$), the operational transition interval $[a, b]$ becomes extremely narrow relative to the total dynamic range of Γ. Consequently, the integration error introduced outside this interval decays exponentially toward zero, rendering this linear approximation mathematically sound, exceptionally tight and highly reliable for asymptotic performance evaluations in short-packet networks.

By exploiting this piecewise linear model, we derive a tractable, closed-form expression for the average BLER.

Theorem 1 Under SPC constraints, the exact average BLER of the FAS link is tightly approximated by:

$$\bar{\epsilon} \approx 1 + \sum_{k=1}^{S} (-1)^k \sum_{\mathcal{N}_k} \frac{\beta\rho}{\Delta_k} \left(e^{-\frac{a\Delta_k}{\rho}} - e^{-\frac{b\Delta_k}{\rho}} \right). \tag{14}$$

Proof: The proof is provided in Appendix A.

3.3 Asymptotic Analysis and Diversity Order

To gain deeper fundamental insights into the system limits, we perform an asymptotic analysis in the high-SNR regime ($\rho \to \infty$).

Theorem 2 In the high-SNR regime ($\rho \to \infty$), the average BLER of the FAS system decays asymptotically according to:

$$\bar{\epsilon}^{\infty} \approx \frac{\beta}{S+1}\left(\frac{1}{\prod_{i=1}^{S} \mu_i}\right)\left(\frac{1}{\rho}\right)^{S} (b^{S+1} - a^{S+1}). \tag{15}$$

Consequently, the diversity order of the FAS system under SPC is exactly equal to S, the rank of the spatial correlation matrix.

Proof: The asymptotic expansion is derived in Appendix B and its mathematically rigorous proof based on exact channel gain bounds and the Squeeze Theorem is detailed in Appendix C.

Remark 1 In a traditional MIMO system with K independent antennas, the diversity order is strictly K. However, due to the severe spatial correlation in FAS, the effective number of independent paths is bounded by S. Even if the number of ports K approaches infinity within a fixed space L_s, the diversity order S will eventually saturate, highlighting the fundamental physical limit of the FAS configuration.

3.4 Effective System Throughput Analysis

Beyond the system reliability evaluated by the average BLER, another crucial performance metric in short-packet communications is the effective system throughput. In the finite-blocklength regime, the throughput is defined as the expected fraction of information bits successfully delivered to the receiver per channel use.

Given that the transmitter sends D information bits over a blocklength of L channel uses, the nominal coding rate is $R_c = D/L$ (bpcu). Taking into account the transmission-failure probability characterized by the average BLER $\bar{\epsilon}$, the effective system throughput, denoted as η, can be formulated as [32]:

$$\eta = R_c(1 - \bar{\epsilon}) = \frac{D}{L}(1 - \bar{\epsilon}). \tag{16}$$

By substituting the exact average BLER derived in Theorem 1 into (16), we obtain the exact closed-form expression for the effective system throughput of the FAS link:

$$\eta \approx R_c \left[1 + \sum_{k=1}^{S} (-1)^k \sum_{\mathcal{N}_k} \frac{\beta\rho}{\Delta_k}\left(e^{-\frac{a\Delta_k}{\rho}} - e^{-\frac{b\Delta_k}{\rho}}\right)\right]. \tag{17}$$

Furthermore, utilizing the asymptotic BLER from Theorem 2, the high-SNR asymptotic throughput can be straightforwardly expressed as:

$$\eta^{\infty} \approx R_c \left[1 - \frac{\beta}{S+1}\left(\frac{1}{\prod_{i=1}^{S} \mu_i}\right)\left(\frac{1}{\rho}\right)^{S} (b^{S+1} - a^{S+1})\right]. \tag{18}$$

Remark 2 It is worth noting that the effective system throughput definition in (16) serves as an optimistic physical-layer performance boundary, as it explicitly omits packet retransmissions and queueing delays. In practical URLLC deployments characterized by bursty traffic arrivals and finite-buffer dynamics, packet-drop policies and automatic repeat request (ARQ) overheads would introduce additional queueing delays and latency violations. Incorporating these finite-buffer cross-layer interactions remains a vital direction for future network-layer extensions.

Remark 3 Equation (18) explicitly demonstrates that in the asymptotically high-SNR regime ($\rho \to \infty$), the average BLER decays to zero ($\bar{\epsilon}^{\infty} \to 0$) and the effective system throughput reaches its maximum saturation limit, $\eta \to R_c$. However, the speed at which the system approaches this maximum limit is strictly governed by the physical diversity order S.

3.5 System Reliability

In the context of short-packet communications, system reliability is defined as the probability that a transmitted finite-length packet is successfully decoded and accurately identified at the receiver. Based on the average BLER evaluated in Theorem 1, the overall system reliability of the FAS link can be mathematically expressed as a percentage [35], Eq. (27):

$$\mathcal{R} = (1 - \bar{\epsilon}) \times 100\%. \tag{19}$$

By substituting the exact closed-form expression of the average BLER from (14) into (19), we can directly quantify the exact reliability of the proposed system. Furthermore, utilizing the high-SNR approximations, the asymptotic reliability readily approaches $\mathcal{R}^{\infty} = (1 - \bar{\epsilon}^{\infty}) \times 100\%$, where $\bar{\epsilon}^{\infty}$ is strictly governed by the physical-diversity order S, as derived in (15).

3.6 Optimal Blocklength Allocation for Throughput Maximization

In short-packet communications, the payload size D is typically predetermined by the application layer, whereas the blocklength L serves as a critical design degree of freedom. A smaller L elevates the coding rate $R_c = D/L$, which triggers a sharp growth in the average BLER$\bar{\epsilon}(L)$, thereby severely degrading the effective system throughput. Conversely, an excessively large L ensures low block errors, but causes a drastic penalty in spectral efficiency due to the over-allocation of time-frequency slots. This fundamental conflict motivates the formulation of an optimization problem to find the optimal blocklength L^* that maximizes the effective system throughput under strict latency and reliability boundaries:

$$\begin{aligned} \max_{L} \eta(L) \quad & = \frac{D}{L}(1 - \bar{\epsilon}(L)) \\ \text{s.t. } L_{\min} \quad & \leq L \leq L_{\max} \\ \bar{\epsilon}(L) & \leq \epsilon_{\max} \end{aligned} \tag{20}$$

where $L_{\min}$ and $L_{\max}$ represent the operational latency limits and $\epsilon_{\max}$ denotes the maximum tolerable block error rate. Due to the complex non-linear coupling of L within both the coding rate and the channel dispersion in the average BLER formula (14), obtaining a closed-form derivative with respect to L is mathematically intractable. However, since the objective function $\eta(L)$ is strictly quasi-concave with respect to L, the unique optimal blocklength L^* can be efficiently and reliably extracted *via* low-complexity numerical search schemes, such as the golden-section search or the standard bisection method.

4. NUMERICAL RESULTS

In this section, we provide extensive numerical results to validate our derived analytical expressions and explore the system's behavior. Unless otherwise specified, we set the normalized physical space to $L_s = 1$. For the short-packet parameters, we assume that the transmitter sends an information payload of $D = 120$ bits over a blocklength of $L = 100$ channel uses, yielding a nominal coding rate of $R_c = 1.2$ bpcu. The average transmit SNR ρ varies from -5 dB to 25 dB and the number of fluid ports is evaluated for $K \in \{2,3,4,30\}$. Furthermore, based on the spatial correlation model in (C.5), we generate 10^7 independent random channel realizations to compute the sample mean for all exhaustive Monte Carlo simulation results.

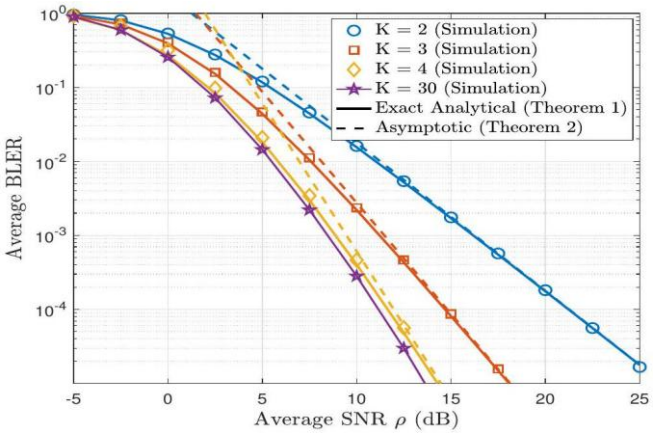


Figure 1. Average BLER *versus* ρ with $L_s = 1$.

In Figure 1, we plot the average BLER *versus* the average transmit SNR ρ for a fixed normalized length $L_s = 1$ with different numbers of fluid ports K. As expected, the average BLER decreases significantly as K increases. The curves exhibit a steeper slope at high SNR regimes, which strictly reflects a higher

diversity order compared to the traditional single-antenna baseline ($K = 1$). However, it is observed that as K grows beyond 4 to a much larger value (e.g., $= 30$), the performance improvement becomes marginal and the curves begin to pack closely together. This behavior perfectly aligns with our theoretical findings in Theorem 2 and the physical intuition discussed in Remark 1. Specifically, the asymptotic analysis mathematically reveals that the diversity order does not scale indefinitely with K, but is fundamentally limited by S, the effective rank of the spatial correlation matrix **R**. For a fixed physical space L_s, packing more ports densely together drastically strengthens the spatial correlation among them. This extreme correlation restricts the rank S and ultimately saturates the achievable diversity order and spatial degrees of freedom for short-packet communications.

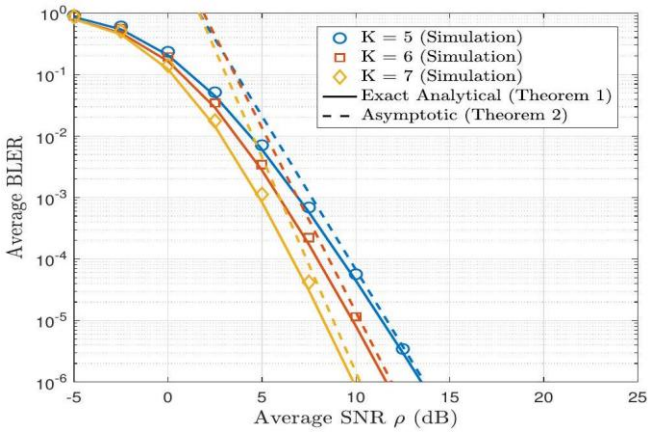


Figure 2. Average BLER *versus* ρ with $L_s = 3$.

To explore the effect of expanding the physical size of the fluid antenna, we plot the average BLER for a larger normalized space of $L_s = 3$ in Figure 2. Compared to the severely constrained space ($L_s = 1$) in Figure 1, expanding the physical dimension allows the K ports to be distributed with significantly reduced spatial correlation. This reduction in correlation inherently increases the effective rank S of the spatial correlation matrix **R**, thereby unlocking a higher effective diversity order. Consequently, as K increases, the average BLER improves even more drastically and each corresponding curve exhibits a noticeably steeper slope. Notably, the performance gap between $K = 4$ and $K = 30$ is much more pronounced here than in Figure 1, indicating that the diversity order has not yet saturated as severely. Furthermore, across both Figure 1 and Figure 2, the exact analytical BLER derived in Theorem 1 (solid lines) matches the exhaustive Monte Carlo simulations (symbols) flawlessly. At the same time, the highSNR asymptotic approximations derived in Theorem 2 (dashed lines) tightly converge with the exact curves in the high-SNR regime. This convergence definitively validates the accuracy of our theoretical derivations and confirms that the diversity order is strictly governed by the rank S.

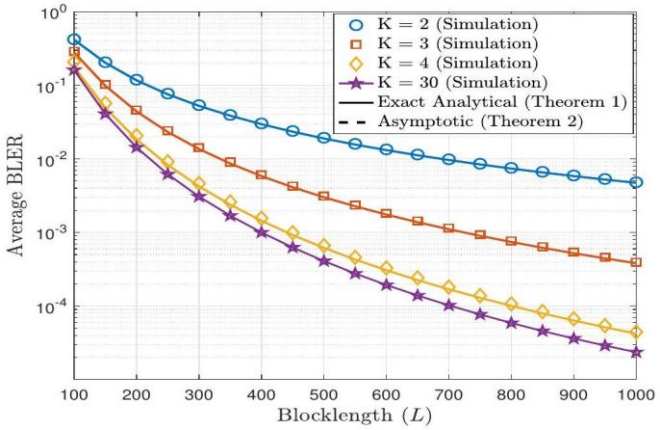


Figure 3. Average BLER *versus* Blocklength L with $L_s = 1$ and $= 1$ dB.

In Figure 3, we investigate the impact of the blocklength L on the average BLER while maintaining a fixed physical space $L_s = 1$ and $K = \{2,3,4,30\}$ ports. As L increases from 100 to 1000 channel uses, the BLER curves shift significantly towards the lower SNR region. This aligns with short-packet

communication principles, where a larger L for a fixed payload D yields a lower effective coding rate R_c, enabling the system to achieve the same reliability target with a lower transmit power.

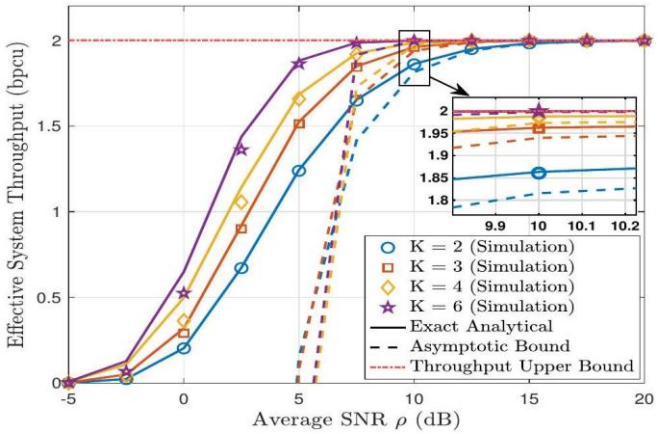


Figure 4. Effective system throughput *versus* ρ with $L_s = 2, D = 200$ bits and $L = 100$.

Figure 4 evaluates the effective system throughput *versus* the average SNR ρ for $L_s = 2, D = 200$ bits and $L = 100$. As expected, the throughput increases monotonically with ρ and eventually saturates at the theoretical upper bound of $R_c = 2.0$ bpcu. Furthermore, deploying a denser FAS configuration (e.g., increasing K to 6) significantly shifts the curves to the left, enabling the system to reach near-maximum throughput at a much lower transmit power. Finally, the exact agreement between the analytical expressions, asymptotic bounds and Monte Carlo simulations firmly validates our throughput derivations in Sub-section 3.3.

Figure 5 illustrates the system reliability *versus* the average SNR ρ for $L_s = 2, D = 200$ bits and $L = 100$. As expected, the reliability monotonically increases with ρ and ultimately saturates at the absolute maximum of 100%. Notably, deploying a denser FAS configuration (e.g., $= 6$) drastically shifts the curves towards the lower SNR region. This demonstrates that a higher number of fluid ports allows the system to guarantee near-perfect packet delivery at a significantly reduced transmit power. Finally, the exact match among the analytical expressions, asymptotic bounds and Monte Carlo simulations rigorously validates our reliability derivations in Sub-section 3.5.

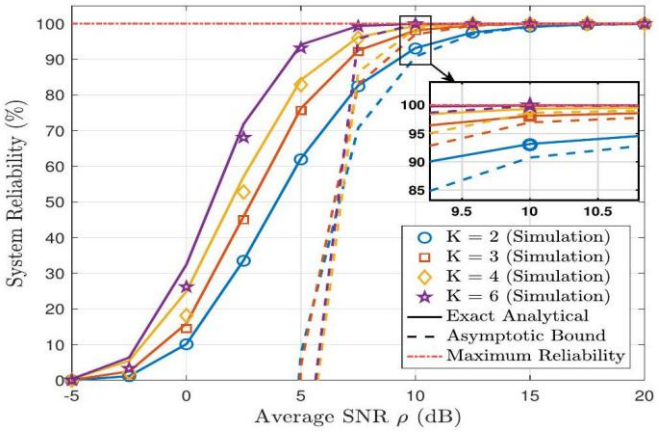


Figure 5. System reliability *versus* ρ with $L_s = 2, D = 200$ bits and $L = 100$.

Figure 8 illustrates the effective throughput η *versus* blocklength L for various numbers of FAS ports $K \in \{2,3,4,30\}$. As observed from the figure, all throughput curves exhibit a distinct single-peak behavior, which numerically validates the quasi-concavity of the objective function with respect to L. When L is relatively small, the throughput drops to zero because the short blocklength cannot combat the severe channel fading, leading to a high BLER. As L increases, the BLER decreases significantly, causing a sharp rise in throughput until it reaches the global maximum at the optimal blocklength L^*, denoted by the red stars. Beyond this optimal point, further increasing L diminishes the transmission rate D/L, thereby resulting in a gradual decline in the effective system throughput. Furthermore, the results highlight the significant advantage of deploying a large number of FAS ports. Specifically,

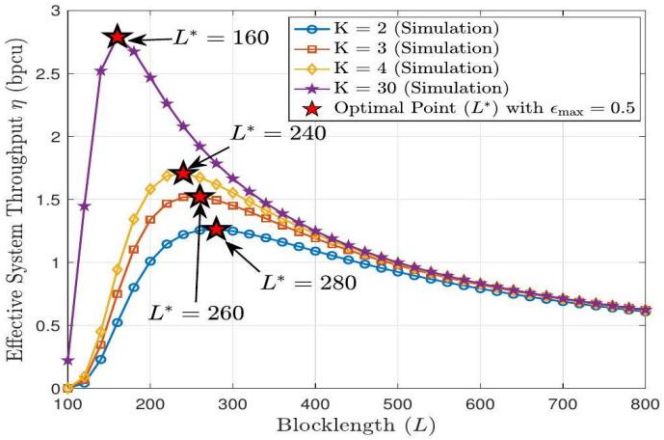


Figure 6. Effective system throughput *versus* blocklength L for $L_s = 3$ and $D = 500$ bits.

increasing K from 2 to 30 drastically enhances the maximum achievable throughput due to the enhanced spatial diversity gain of fluid antennas. Crucially, a larger K shifts the optimal blocklength L^* toward a smaller value (i.e., to the left of the horizontal axis). This indicates that by exploiting the fine-grained port selection in FAS, the system can achieve peak throughput at a much shorter blocklength, effectively minimizing the transmission latency while maintaining high reliability, which is highly desirable for short-packet URLLC systems.

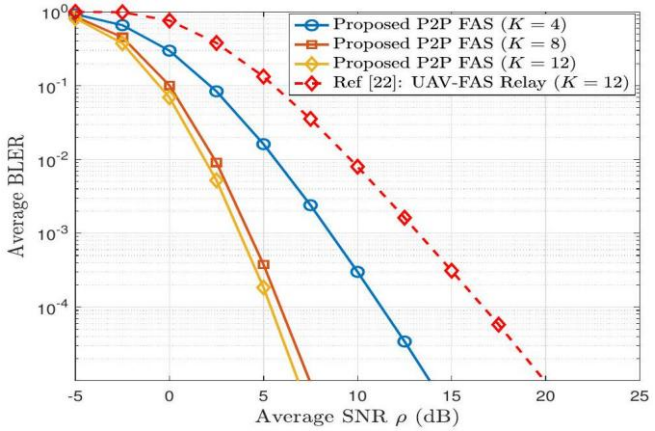


Figure 7. Average BLER *versus* ρ for $L_s = 3, D = 120$ bits and $L = 100$ under the optimistic constant correlation model [23].

Figure 7 illustrates the average BLER *versus* the average SNR ρ for the proposed P2P FAS ($K \in \{3,8,12\}$) alongside the UAV-FAS relay benchmark [23]. As observed, increasing the number of fluid ports from $K = 3$ to $K = 12$ drastically enhances system reliability, shifting the BLER curves to the lower-SNR region.

Crucially, under the same configuration (e.g., $= 12$), our proposed direct architecture significantly outperforms the UAV relay benchmark [23]. The relay topology suffers from a severe half-duplex penalty, which forces a doubled coding rate and heavily degrades decoding accuracy under strict short-packet constraints. By completely bypassing this penalty, our P2P FAS framework proves to be a vastly superior and power-efficient solution for practical URLLC applications.

Finally, Figure 8 illustrates the average BLER performance *versus* the average transmit SNR ρ for both the proposed high-density FAS and the traditional ideal selection combining (SC) schemes under SPC constraints. To establish a realistic benchmark, the spatial dimension is strictly constrained to $L_s = 1$. Within this normalized limited space, traditional SC can deploy at most $K = 3$ co-located antennas to maintain independent fading channel characteristics. As observed from the simulation results, the proposed FAS with $K = 30$ ports significantly outperforms the physical boundaries of the traditional ideal SC ($= 2$ and $K = 3$). Specifically, at a target BLER of 10^{-4}, the proposed FAS achieves approximately a 3 dB SNR gain compared to the maximum bound of the traditional ideal SC($K = 3$). Furthermore, the curves clearly indicate that FAS provides a sharper slope in the high-SNR regime. This

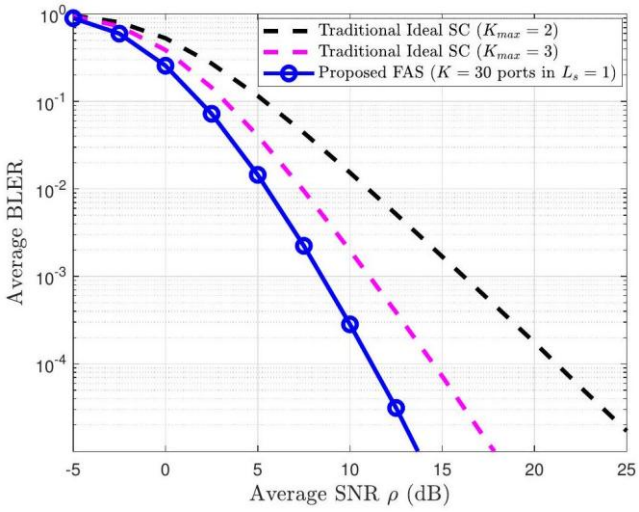


Figure 8. Average BLER *versus* average transmit SNR ρ for the proposed FAS ($K = 30$) and traditional ideal SC ($K = 2,3$) under short-packet transmission constraints ($L_s = 1, D = 120$ bits and $L_{\text{block}} = 100$).

behavior is attributed to the fact that although the ultra-dense port arrangement in FAS introduces severe spatial correlation, the effective rank of the Jakes' correlation matrix over $L_s = 1$ is approximately 4. By deploying $K = 30$ ports, FAS successfully captures all available spatial degrees of freedom and maximizes the array gain through fine-grained selection, whereas traditional SC is fundamentally restricted by hardware and spatial constraints. These findings firmly highlight the superiority of fluid-antenna technology in breaking the physical limits of conventional multi-antenna systems for ultra-reliable low-latency IoT networks.

5. CONCLUSION

In this paper, we rigorously analyzed the delivery performance of FAS under the strict and highly realistic constraints of short-packet communications, departing from the conventional infinite-blocklength Shannon capacity framework. By accurately modeling the heavily correlated spatial fading channels *via* eigenvalue decomposition, we successfully derived highly accurate, closed-form expressions for the exact average BLER, the effective system throughput and the overall system reliability. Furthermore, our high-SNR asymptotic analysis mathematically proved that the diversity order of an FAS link is strictly determined and bounded by the rank of its spatial correlation matrix, rather than the raw number of fluid ports. The comprehensive numerical results flawlessly confirmed our theoretical derivations, highlighting that densely packed FAS configurations can achieve near-perfect reliability and maximum throughput at significantly reduced transmit power levels. Ultimately, these findings demonstrate that FAS is a highly promising architectural solution for meeting the extreme delivery demands of modern IoT and URLLC networks, offering substantial performance leaps without the need for bulky, power-hungry multiple-RFchain hardware typical of traditional multi-antenna setups.

APPENDICES

A. PROOF OF THEOREM 1

By substituting the piecewise linear approximation (10) into the expectation integral (9), we obtain:

$$\bar{\epsilon} \approx \int_0^a 1 \cdot f_\Gamma(x)dx + \int_a^b \left[\frac{1}{2} - \beta(x-\theta)\right] f_\Gamma(x)dx. \tag{A.1}$$

The first term trivially evaluates to $F_\Gamma(a)$. For the second term, we apply integration by parts. Let $u = \frac{1}{2} - \beta(x-\theta)$, yielding $du = -\beta dx$. Let $dv = f_\Gamma(x)dx$, yielding $v = F_\Gamma(x)$. Evaluating the boundaries gives:

$$\left[\frac{1}{2} - \beta(x-\theta)\right] F_\Gamma(x)\Big|_a^b + \beta \int_a^b F_\Gamma(x)dx. \tag{A.2}$$

Due to the symmetric definitions of a and b around θ in (12), the boundary evaluations exactly cancel out the $F_\Gamma(a)$ term. The average BLER simplifies to:

$$\bar{\epsilon} \approx \beta \int_a^b F_\Gamma(x)dx. \tag{A.3}$$

Optionally, substituting the CDF expression from (6) into (A.3) yields:

$$\bar{\epsilon} \approx \beta \int_a^b \left(1 + \sum_{k=1}^{S} (-1)^k \sum_{\mathcal{N}_k} e^{-\frac{x\Delta_k}{\rho}}\right) dx. \tag{A.4}$$

Evaluating the integral of the exponential function, we have:

$$\begin{aligned}\bar{\epsilon} =& \approx \beta(b-a) + \sum_{k=1}^{S} (-1)^k \sum_{\mathcal{N}_k} \beta \int_a^b e^{-\frac{x\Delta_k}{\rho}} dx \\ &= 1 + \sum_{k=1}^{S} (-1)^k \sum_{\mathcal{N}_k} \frac{\beta\rho}{\Delta_k}\left(e^{-\frac{a\Delta_k}{\rho}} - e^{-\frac{b\Delta_k}{\rho}}\right).\end{aligned} \tag{A.5}$$

Note that $\beta(b-a) = 1$ is used in the final step. This completes the proof.

B. PROOF OF THEOREM 2

To uncover the asymptotic behavior ($\rho \to \infty$), we simplify the CDF$F_\Gamma(x)$. The exact CDF before applying the inclusion-exclusion principle is a product of independent exponential variables:

$$F_\Gamma(x) = \prod_{i=1}^{S} \left(1 - e^{-\frac{x}{\rho\mu_i}}\right). \tag{B.1}$$

As $\rho \to \infty$, the term $\frac{x}{\rho\mu_i} \to 0$. Applying the first-order Maclaurin series expansion $e^{-z} \approx 1 - z$, the CDF simplifies to:

$$F_\Gamma^{\infty}(x) \approx \prod_{i=1}^{S} \left(\frac{x}{\rho\mu_i}\right) = \left(\frac{1}{\prod_{i=1}^{S} \mu_i}\right)\left(\frac{1}{\rho}\right)^S x^S. \tag{B.2}$$

Substituting (B.2) into the simplified integral (A.3) gives:

$$\begin{aligned}\bar{\epsilon}^{\infty} &\approx \beta \int_a^b F_\Gamma^{\infty}(x)dx = \beta\left(\frac{1}{\prod_{i=1}^{S} \mu_i}\right)\left(\frac{1}{\rho}\right)^S \int_a^b x^S dx \\ &= \frac{\beta}{S+1}\left(\frac{1}{\prod_{i=1}^{S} \mu_i}\right)\left(\frac{1}{\rho}\right)^S (b^{S+1} - a^{S+1}).\end{aligned} \tag{B.3}$$

This completes the proof.

C. RIGOROUS PROOF OF DIVERSITY ORDER *VIA* CHANNEL BOUNDS

To establish the exact diversity order without relying on any distribution approximation of the SNR, we bound the true combined channel gain $|h_{\mathrm{FAS}}|^2 \triangleq \max_{1\le k\le K} |h_k|^2$. *Via* the eigenvalue decomposition of the spatial correlation matrix $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^H$, the channel vector can be modeled as $\mathbf{h} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{z}$, where $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_K)$. The squared norm of the channel vector satisfies $\|\mathbf{h}\|^2 = \sum_{m=1}^{S} \mu_m |z_m|^2$, where $S = \mathrm{Rank}(\mathbf{R})$ and μ_m represents the ordered non-zero eigenvalues. Following the bounding methodology in [29], we utilize the inequality $\frac{1}{K}\|\mathbf{h}\|^2 \le |h_{\mathrm{FAS}}|^2 \le \|\mathbf{h}\|^2$, which yields:

$$\frac{\mu_{\min}}{K} Z \le |h_{\mathrm{FAS}}|^2 \le \mu_{\max} Z, \tag{C.1}$$

where $Z = \sum_{m=1}^{S} |z_m|^2$ is a Gamma-distributed random variable with shape parameter S and scale parameter 1 , while $\mu_{\min}$ and $\mu_{\max}$ denote the minimum and maximum non-zero eigenvalues of $\mathbf{R}$, respectively.

Let $\Gamma = \rho|h_{\mathrm{FAS}}|^2$ be the exact instantaneous SNR. Its true CDF, $F_\Gamma(x)$, is strictly bounded by $F_\Gamma^{\text{lower}}(x) \le F_\Gamma(x) \le F_\Gamma^{\text{upper}}(x)$, where $F_\Gamma^{\text{lower}}(x) = \Pr\left(Z \le \frac{x}{\mu_{\max}\rho}\right)$ and $F_\Gamma^{\text{upper}}(x) = \Pr\left(Z \le \frac{Kx}{\mu_{\min}\rho}\right)$.

By applying integration by parts as detailed in Appendix A, the exact average BLER $\bar{\epsilon}$ under short-packet constraints is expressed as $\bar{\epsilon} = \beta \int_a^b F_\Gamma(x)dx$. Since the operation preserves the inequality directions, we obtain $\bar{\epsilon}^{\text{lower}} \le \bar{\epsilon} \le \bar{\epsilon}^{\text{upper}}$, where $\bar{\epsilon}^{\text{lower}} = \beta \int_a^b F_\Gamma^{\text{lower}}(x)dx$ and $\bar{\epsilon}^{\text{upper}} = \beta \int_a^b F_\Gamma^{\text{upper}}(x)dx$.

In the high-SNR regime ($\rho \to \infty$), the standard Gamma CDF near zero expands asymptotically *via* Maclaurin series as $F_\Gamma^{\text{lower}}(x) \simeq \frac{1}{S!}\left(\frac{x}{\mu_{\max}\rho}\right)^S$ and $F_\Gamma^{\text{upper}}(x) \simeq \frac{1}{S!}\left(\frac{Kx}{\mu_{\min}\rho}\right)^S$. Substituting these polynomial expansions back into the integrals, both the upper and lower bounds of the average BLER converge to the identical mathematical order of $\mathcal{O}(\rho^{-S})$. Finally, invoking the Squeeze Theorem, the exact diversity order d is rigorously established as:

$$d = -\lim_{\rho\to\infty} \frac{\ln \bar{\epsilon}}{\ln \rho} = S = \mathrm{Rank}(\mathbf{R}). \tag{C.2}$$

This completes the mathematically rigorous proof and demonstrates that the asymptotic behavior is unaltered by the distribution approximation.

REFERENCES

[1] B. Hassan, S. Baig and M. Asif, "Key Technologies for Ultra-reliable and Low-latency Communication in 6G," IEEE Communications Standards Magazine, vol. 5, no. 2, pp. 106-113, 2021.

[2] M. Adhikari and A. Hazra, "6G-enabled Ultra-reliable Low-latency Communication in Edge Networks," IEEE Communications Standards Magazine, vol. 6, no. 1, pp. 67-74, 2022.

[3] T. Q. Duong, D. Van Huynh, S. R. Khosravirad, V. Sharma, O. A. Dobre and H. Shin, "From Digital Twin to Metaverse: The Role of 6G Ultra-reliable and Low-latency Communications with Multitier Computing," IEEE Wireless Communications, vol. 30, no. 3, pp. 140-146, 2023.

[4] X. Zhao, J. Liu, Y. Wang and J. Wang, "CryptoMamba-SSM: Linear Complexity State Space Models for Cryptocurrency Volatility Prediction," IEEE Open J. of the Computer Society, vol. 7, pp. 226-243, 2026.

[5] Y. Zhu, Y. Hu, X. Yuan, M. C. Gursoy, H. V. Poor and A. Schmeink, "Joint Convexity of Error Probability in Blocklength and Transmit Power in the Finite Blocklength Regime," IEEE Transactions on Wireless Communications, vol. 22, no. 4, pp. 2409-2423, 2023.

[6] M. Yin, Y. Yang, J.-M. Wu and B. Jiao, "Opportunistic Bits in Short-packet Communications: A Finite Blocklength Perspective," IEEE Transactions on Communications, vol. 69, no. 12, pp. 80858099, 2021.

[7] C. Feng, H.-M. Wang and H. V. Poor, "Reliable and Secure Short-packet Communications," IEEE Transactions on Wireless Communications, vol. 21, no. 3, pp. 1913-1926, 2022.

[8] C. M. W. Basnayaka, D. N. K. Jayakody, T. D. P. Perera and M. Beko, "DataAge: Age of Information in SWIPT-driven Short Packet IoT Wireless Communications," IEEE Internet of Things Journal, vol. 11, no. 16, pp. 26984-26999, 2024.

[9] F. A. Pereira de Figueiredo, "An Overview of Massive MIMO for 5G and 6G," IEEE Latin America Transactions, vol. 20, no. 6, pp. 931-940, 2022.

[10] H. Q. Ngo, G. Interdonato, E. G. Larsson, G. Caire and J. G. Andrews, "Ultradense Cell-free Massive MIMO for 6G: Technical Overview and Open Questions," Proceedings of the IEEE, vol. 112, no. 7, pp. 805-831, 2024.

[11] N. T. Nguyen, V.-D. Nguyen, H. V. Nguyen, H. Q. Ngo, A. L. Swindlehurst and M. Juntti, "Performance Analysis and Power Allocation for Massive MIMO ISAC Systems," IEEE Transactions on Signal Processing, vol. 73, pp. 1691-1707, 2025.

[12] Y. Lu, J. Zhang, E. Shi, P. Zhang, D. W. Kwan Ng, D. Niyato and B. Ai, "Energy-efficient RIS-Aided Cell-free Massive MIMO Systems: Application, Opportunities and Challenges," IEEE Wireless Communications, vol. 32, no. 4, pp. 148-155, 2025.

[13] B. Di, H. Zhang, Z. Han, R. Zhang and L. Song, "Reconfigurable Holographic Surface: A New Paradigm for Ultra-massive MIMO," IEEE Transactions on Cognitive Communications and Networking, vol. 11, no. 6, pp. 3761-3783, 2025.

[14] T. C. Rapudu and O. O. Oyerinde, "Machine Learning-based Channel Estimation for Multi-RIS-assisted mmWave Massive-MIMO OFDM System in a Dynamic Environment," IEEE Transactions on Wireless Communications, vol. 24, no. 6, pp. 5297-5309, 2025.

[15] K. N. Le, "Impact of Fading Correlation and Finite-blocklength Regimes on Secrecy," IEEE Transactions on Vehicular Technology, vol. 72, no. 3, pp. 4068-4072, 2023.

[16] C. Psomas, P. J. Smith, H. A. Suraweera and I. Krikidis, "Continuous Fluid Antenna Systems: Modeling and Analysis," IEEE Communications Letters, vol. 27, no. 12, pp. 3370-3374, 2023.

[17] C. Wang, Z. Li, K.-K. Wong, R. Murch, C.-B. Chae and S. Jin, "Ai-empowered Fluid Antenna Systems: Opportunities, Challenges and Future Directions," IEEE Wireless Communications, vol. 31, no. 5, pp. 34-41, 2024.

[18] W. K. New, K.-K. Wong, C. Wang, C.-B. Chae, R. Murch, H. Jafarkhani and Y. Hao, "Fluid Antenna Systems: Redefining Reconfigurable Wireless Communications," IEEE Journal on Selected Areas in Communications, vol. 44, pp. 1013-1044, 2026.

[19] T. Gong, P. Gavriilidis, R. Ji, C. Huang, G. C. Alexandropoulos, L. Wei, Z. Zhang, M. Debbah, H. V. Poor and C. Yuen, "Holographic MIMO Communications: Theoretical Foundations, Enabling Technologies and Future Directions," IEEE Communications Surveys & Tutorials, vol. 26, no. 1, pp. 196-257, 2024.

[20] K.-F. Tong, B. Liu and K.-K. Wong, "Designs and Challenges in Fluid Antenna System Hardware," Electronics, vol. 14, no. 7, [Online], Available: https://www.mdpi.com/2079-9292/14/7/1458, 2025.

[21] D. Bepari, S. Mondal, A. Bhowal, K. Singh, H. Shin, K.-K. Wong and D. W. K. Ng, "Fluid Antennas: Pioneering a New Era in Next-generation Wireless Communications," IEEE Communications Surveys & Tutorials, vol. 28, pp. 5596-5631, 2026.

[22] P. Ramírez-Espinosa et al., "A New Spatial Block-correlation Model for Fluid Antenna Systems," IEEE Trans. on Wireless Communications, vol. 23, no. 11, pp. 15829-15843, 2024.

[23] X. Zhu et al., "UAV-enabled Short-packet Communication *via* Fluid Antenna Systems," arXiv: 2602.06206, DOI: 10.48550/arXiv.2602.06206, 2026.

[24] Z. Zhang et al., "Finite-blocklength Fluid Antenna Systems," arXiv: 2509.15643**,** [Online], Available: https://arxiv.org/abs/2509.15643, 2026.

[25] X. Zhu et al., "Fluid Antenna Systems: A Geometric Approach to Error Probability and Fundamental Limits," IEEE Trans. on Wireless Communications, vol. 25, pp. 17195-17209, 2026.

[26] X. Zhu et al., "Fluid Antenna Systems Enabling 6G HRLLC with Port Switching Delay," arXiv: 2605.06275, [Online], Available: https://arxiv.org/abs/2605.06275, 2026.

[27] X. Zhu et al., "Fluid Antenna System-enabled UAV Communications in the Finite Blocklength Regime," IEEE Transactions on Wireless Communications, vol. 25, pp. 16714-16729, 2026.

[28] G. L. Stüber, Principles of Mobile Communication, 4th Edn, ISBN: 978-3-319-55615-4, Springer Nature, 2017.

[29] H. Zhao and D. Slock, "Analytical Insights into Outage Probability and Ergodic Capacity of Fluid Antenna Systems," IEEE Wireless Communications Letters, vol. 14, no. 5, pp. 1581-1585, 2025.

[30] M. Khammassi et al., "A New Analytical Approximation of the Fluid Antenna System Channel," IEEE Trans. on Wireless Communications, vol. 22, no. 12, pp. 8843-8858, 2023.
[31] W. K. New et al., "Fluid Antenna System: New Insights on Outage Probability and Diversity Gain," IEEE Trans. on Wireless Communications, vol. 23, no. 1, pp. 128-140, 2024.
[32] T.-H. Vu, T.-V. Nguyen, Q.-V. Pham, D. Benevides da Costa and S. Kim, "STAR-RIS-enabled Short-packet NOMA Systems," IEEE Trans. on Vehicular Technology, vol. 72, no. 10, pp. 13764-13769, 2023.
[33] T.-T. T. Nguyen, X.-X. Nguyen and N. V. T. Nguyen, "Short-packet Communications for Pinching-antenna Systems," IEEE Communications Letters, vol. 30, pp. 1270-1274, 2026.
[34] L. T. T. Huyen, B. A. Duc, X. N. Tran, B. C. Nguyen, L. T. Dung and T. M. Hoang, "BLER and BER Performance of RSMA in Multi-antenna Systems over Nakagami Fading Channels," IEEE Internet of Things Journal, Early Access, DOI: 10.1109/JIOT.2026.3686852, 2026.
[35] G. N. Tran and S. Kim, "Performance Analysis of Short Packets in NOMA VLC Systems," IEEE Access, vol. 10, pp. 6505-6517, 2022.

ملخص البحث:

توفّر أنظمة الهوائيات المائعة تنوُّعاً مكانياً هائلاً ضمن مساحاتٍ مُدمجة. ومع ذلك، فإنّ تحليلات الأداء الحالية تعتمد في الغالب على حُدود سِعة "شانون"، بافتراض إرسال كُتل بياناتٍ ذات طولٍ لا نهائي. ولتلبية المتطلّبات الصّارمة لتطبيقات إنترنت الأشياء، تبحث هذه الورقة في نظام هوائياتٍ مائعة يعمل بنمط الاتّصال من نقطةٍ إلى نقطةٍ في ظلّ قيود الاتّصالات ذات الحُزم القصيرة.

يتمّ اشتقاق دقيق لِصيغ تحليلية مغلقة تعبّر عن متوسّط معدّل خطأ الكتلة، والإنتاجية الفعلية للنّظام، والموثوقية. كذلك نجري تحليلاً تقاربياً في نطاق نسبة الإشارة إلى الضّوضاء العالية، ونكشف رياضياً أنّ رُتبة التّنوّع في النّظام تتحدّد بشكلٍ صارمٍ بناءً على رتبة مصفوفة الارتباط المكاني.

وتبيّن محاكاة مونت كارلو الدّقة العالية للنّماذج التّحليلية، حيث تؤدي زيادة عدد المنافذ المائعة إلى خفض معدّل خطأ الكتلة بشكلٍ ملحوظ.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

ADVANCEMENTS IN AI-DRIVEN DIABETIC RETINOPATHY DETECTION: A SYSTEMATIC SURVEY OF METHODS, CHALLENGES AND FUTURE DIRECTIONS

Yasser A. Ali[1], Ahmed Z. Emam[2], Muhammed H. Zayyan[1], Samir Elmougy[1] and Ahmed S. Tolba[1]

(Received: 25-Mar.-2026, Revised: 31-May-2026 and 22-Jun.-2026, Accepted: 30-Jun.-2026)

ABSTRACT

Diabetic retinopathy (DR) is a leading cause of preventable vision impairment, requiring scalable and reliable screening solutions. Artificial intelligence (AI), particularly machine learning and deep learning, has significantly advanced the automation of DR classification from retinal fundus images. This paper presents a systematic review of AI-based DR classification approaches, focusing on binary classification, multi-class severity grading and lesion-based methods and introduces a taxonomy based on their clinical objectives. The review covers state-of-the-art models, including convolutional neural networks, vision transformers, attention mechanisms and ensemble techniques, as well as their pre-processing strategies and evaluation protocols. Common benchmark datasets are analyzed in terms of class distribution and their impact on performance. While current models achieve high accuracy on standard datasets, challenges remain in generalization, class imbalance and reliable detection of early-stage DR. Lesion-based approaches improve interpretability, but are often not fully integrated into grading frameworks. Overall, this review highlights key limitations and outlines future directions toward developing robust, interpretable and clinically deployable DR classification systems for real-world screening.

KEYWORDS

Diabetic Retinopathy, Automated Disease Screening, Deep Learning, Artificial Intelligence, Fundus Imaging, Explainable AI, Vision Transformers, Domain Adaptation, Class Imbalance.

1. INTRODUCTION

Diabetic Retinopathy (DR) is a significant public health issue impacting millions globally, resulting in vision impairment and blindness if not addressed, primarily affecting those with chronic diabetes [1],[2],[3]. The illness involves damage to the retinal blood vessels caused by persistent hyperglycemia, which triggers vascular leakage and capillary occlusion [4]-[5]. This process leads to specific abnormalities, including microaneurysms, hemorrhages, exudates and neovascularization [1], [6]. The quick and accurate diagnosis of DR is essential to avert vision impairment, as timely intervention can markedly enhance patient prognoses and prevent the disease from escalating to irreversible stages [2], [4], [7]-[8].

Traditional DR screening relies on manual examination of retinal fundus images by trained ophthalmologists [4], [7], [9]-[10]. Although effective, this approach is time-consuming, costly and subject to inter-observer variability, meaning that the level of agreement among specialists can vary significantly [11]-[12]. Furthermore, many regions—particularly rural and low-resource areas—lack sufficient ophthalmology specialists, leading to delayed diagnosis and treatment [6], [9]. Recent advances in artificial intelligence (AI), particularly machine learning (ML) and deep learning (DL), have transformed medical image analysis and offered promising solutions for automated DR detection [6], [9], [13]. AI-based systems can analyze retinal images efficiently, detect pathological features, classify disease severity and assist clinicians in large-scale screening programs [3], [14]. The availability of large datasets (such as EyePACS and APTOS 2019), improvements in computational power and the success of convolutional neural networks (CNNs) have accelerated research in this domain [15]-[16].

This paper presents a comprehensive review of AI techniques for the automatic detection and classification of DR. It highlights key factors that affect the performance of DR classification systems,

1. Y. A. Ali, M. H. Zayyan, Samir Elmougy and Ahmed S. Tolba are with Department of Computer Science, Faculty of Computers and Information, Mansoura University, Mansoura, Egypt. Emails: yasaliali@std.mans.edu.eg, {mhaggag, mougy, astolba}@mans.edu.eg
2. A. Z. Emam is with Department of Math & Computer Science, Faculty of Science, Menoufiya University, Menoufiya, Egypt. Email: Ahmed.emam@science.menofia.edu.eg

including image quality, feature representation, learning strategies and dataset characteristics. The reviewed studies emphasize the importance of robust AI models for early diagnosis, large-scale screening and clinical decision support. In addition, this paper summarizes current limitations, open challenges and future research directions in DR classification.

Unlike previous surveys that primarily summarize CNN-based DR classification models, this review offers a broader, more critical analysis of recent AI-driven DR detection approaches. The proposed survey introduces a unified taxonomy encompassing binary classification, multi-class grading, hybrid frameworks and lesion-based methods. Furthermore, this work critically examines emerging trends, including vision transformers (ViTs), graph neural networks, multi-modal learning and quantum-inspired approaches, while emphasizing challenges related to dataset bias, cross-dataset generalization, explainability, reproducibility and real-world clinical deployment. The review also highlights the gap between high benchmark performance and practical clinical applicability, thereby providing a more comprehensive perspective on future-research directions.

The remainder of this paper is organized as follows. Section 2 presents the systematic review methodology. Section 3 describes the general framework for automated DR classification. Section 4 presents a taxonomy of classification approaches, including binary and multi-class severity grading. Section 5 reviews existing classification methods. Section 6 summarizes the benchmark datasets commonly used for DR detection. Section 7 presents a critical discussion of the reviewed literature, highlighting methodological challenges, dataset bias, reproducibility concerns, clinical deployment barriers and emerging research directions. Finally, Section 8 concludes the paper by outlining future opportunities to advance robust, interpretable and clinically deployable DR detection systems.

2. SYSTEMATIC REVIEW METHODOLOGY

This review was conducted following a systematic literature review methodology to ensure transparency, reproducibility and comprehensive coverage of recent advances in AI-driven DR detection. Relevant studies published between 2018 and 2025 were retrieved from major scientific databases, including IEEE Xplore, Scopus, Web of Science, PubMed, ScienceDirect and SpringerLink. Search queries included combinations of keywords, such as “DR”, “DL”, “AI”, “retinal fundus images”, “CNN”, “ViT” and “DR grading”.

The inclusion criteria considered peer-reviewed studies that focused on AI-based DR detection or grading using retinal fundus images and reported quantitative evaluation metrics. Studies unrelated to DR classification, non-English publications, duplicate papers and articles lacking sufficient experimental details were excluded. The screening process involved title and abstract filtering, followed by full-text assessment. Duplicate studies were removed and the selected articles were analyzed for methodology, datasets, evaluation metrics, strengths and limitations. A PRISMA-style workflow illustrating the study selection procedure is shown in Figure 1.

2.1 Quality Assessment of Included Studies

To improve the scientific rigor of this systematic review, the included studies were qualitatively assessed rather than ranked using a numerical scoring system. The assessment considered several methodological dimensions, including dataset transparency and characteristics, annotation quality and grading procedures, the suitability of reported evaluation metrics and the handling of class imbalance. Attention was given to the validation strategy, especially the presence of external validation, the level of methodological detail provided to support reproducibility and the use of interpretability techniques, such as explainable AI methods or lesion-based analysis. In addition, the studies were evaluated with respect to their discussion of clinical relevance, practical deployment considerations and acknowledgment of limitations. This qualitative assessment supported the critical synthesis presented throughout the manuscript and facilitated a more balanced interpretation of reported performance beyond accuracy values alone.

3. THEORETICAL FRAMEWORK FOR AUTOMATED DR DETECTION

The general framework for detecting DR comprises sequential stages that transform raw retinal images into clinically meaningful diagnostic outcomes [2], [7], [17]. The process begins with data acquisition,

in which retinal fundus images are captured using specialized imaging devices, such as digital fundus cameras, either in clinical settings or from publicly available datasets, such as EyePACS, Messidor or IDRiD [2], [4], [18]. These images often vary in resolution, illumination and quality, which necessitates the next stage, pre-processing [4], [19]. Pre-processing aims to enhance image quality and standardize the input data by applying techniques, such as resizing, normalization, contrast enhancement (e.g., CLAHE), noise reduction and illumination correction [2], [19]. This stage improves the visibility of critical retinal structures and pathological signs associated with DR, such as blood vessels and lesions [1], [19].

After pre-processing, feature extraction is performed to get discriminative information from the retinal images [2], [4], [20]-[21]. Features may be handcrafted—such as texture, color and morphological descriptors—or automatically learned using DL models, which can capture complex spatial patterns related to DR lesions [1], [4], [22]. Finally, the extracted features are utilized in the classification or segmentation stage, where ML or DL models assess the presence and severity of DR or to localize specific lesions, such as microaneurysms, hemorrhages and exudates [2], [4], [7], [23]. These stages together form an integrated framework that enables accurate, efficient and automated detection of DR, supporting early diagnosis and large-scale screening programs [2], [4], [24], as shown in Figure 2.

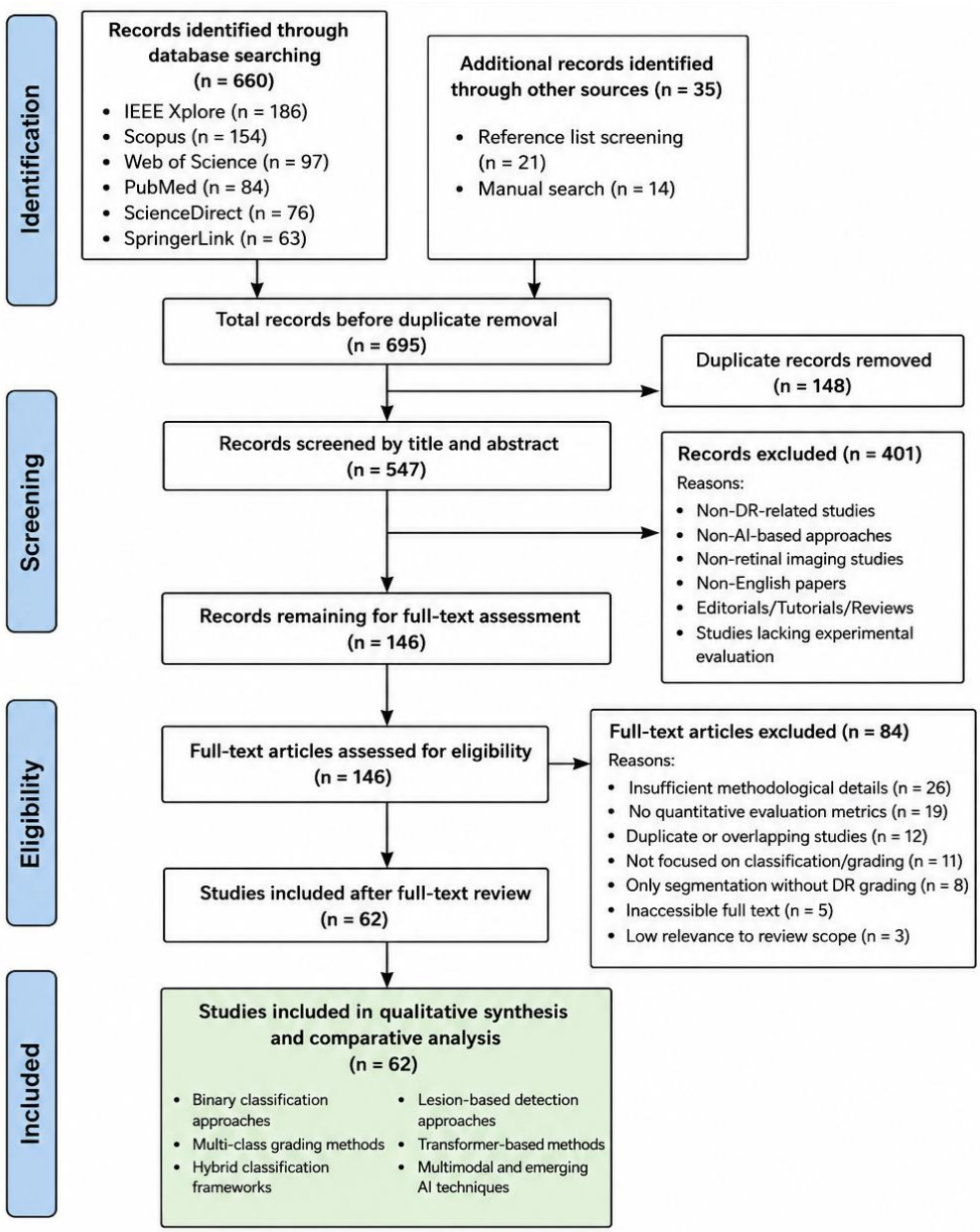


Figure 1. PRISMA-based workflow for identification, screening, eligibility assessment and inclusion of studies in the systematic review.

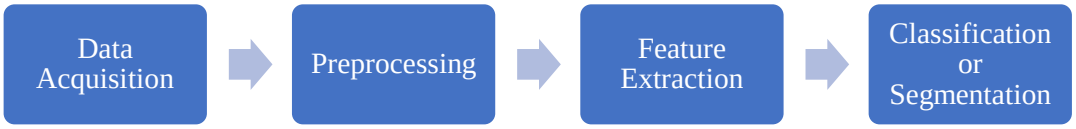


Figure 2. General framework for automated DR detection.

Although lesion segmentation plays an important role in retinal-image analysis by localizing pathological structures, such as microaneurysms, hemorrhages and exudates, the primary focus of this survey is classification-based DR detection and grading approaches. Accordingly, segmentation methods are discussed only in relation to their contribution to feature representation, interpretability and lesion-aware classification frameworks rather than as standalone segmentation systems. A comprehensive analysis of retinal lesion segmentation techniques remains an important direction for future review studies.

4. Taxonomy of DR Detection Approaches

The taxonomy of DR detection approaches, illustrated in Figure 3, is developed through a systematic analysis of the reviewed literature and reflects the major methodological and clinical directions in AI-driven DR diagnosis. The taxonomy categorizes DR detection approaches into four main groups: binary classification, multi-class severity grading, hybrid binary–multi-class frameworks and lesion-based classification approaches. Binary-classification approaches primarily focus on efficient screening tasks, such as distinguishing DR from non-DR or referable from non-referable cases, making them suitable for large-scale screening systems. In contrast, multi-class grading approaches aim to provide more clinically meaningful assessments of severity by categorizing retinal images into multiple DR stages. However, increasing grading granularity introduces greater challenges related to class imbalance, annotation variability and inter-class similarity.

Lesion-based approaches emphasize interpretability by explicitly identifying pathological retinal features, such as microaneurysms, hemorrhages and exudates. These methods provide clinically explainable predictions, but often require detailed lesion annotations and higher computational complexity. More recent hybrid and multi-modal frameworks attempt to combine the strengths of different approaches by integrating lesion localization, global contextual learning and attention-based feature extraction to improve robustness and generalization. The taxonomy also highlights the evolution of DR detection research from performance-oriented classification systems toward more explainable, generalized and clinically deployable AI frameworks. This progression reflects the growing emphasis on interpretability, fairness, cross-dataset generalization and real-world clinical applicability in modern DR detection systems.

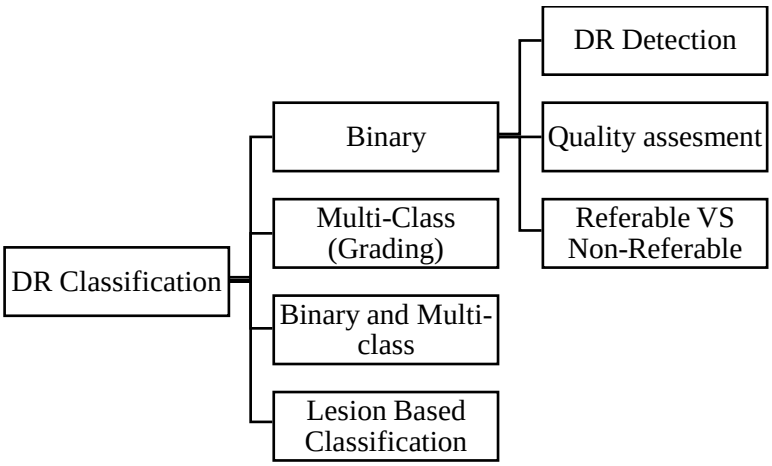


Figure 3. Taxonomy of DR detection.

Based on this taxonomy, the following sections systematically review the literature within these categories. For each category, studies are analyzed in terms of their objectives, methodologies, datasets and evaluation metrics. A critical comparison is also provided, highlighting strengths, limitations and

challenges in terms of accuracy, generalization and clinical applicability. Accordingly, Table 2 summarizes binary classification studies, Table 3 reviews multi-class severity grading, Table 4 presents combined approaches and Table 5 covers lesion-informed classification methods.

5. DR CLASSIFICATION-BASED APPROACHES

The classification of DR is crucial for early diagnosis and treatment, as it ranges from mild non-proliferative stages to severe proliferative forms that may lead to vision loss. Accurate classification helps determine the appropriate management strategies and monitor disease progression, making it a vital aspect of diabetic care [25]. DR is categorized into two main phases: non-proliferative DR (NPDR) and proliferative DR (PDR) [4]. NPDR is further divided into mild, moderate and severe stages based on the presence of specific retinal lesions. The standardized DR severity scale is summarized in Table 1, which outlines 5 distinct stages determined by lesion characteristics observed in retinal fundus images. They can also be categorized as non-referable DR (R0 and R1) and referable DR (R2, R3 and R4).

Table 1. DR grade based on the international clinical DR disease severity scale (ICDRDSS) [26].

Grade	**Description**
R0 - No DR	No visible sign of abnormalities
R1 - Mild NPDR	Only the presence of microaneurysms
R2 - Moderate NPDR	Below severe NPDR and above just microaneurysms
R3 - Severe NPDR	No signs of PDR and any of the following: 1. Each of the four quadrants has above 20 intraretinal hemorrhages 2. Two or more quadrants have definite venous beading 3. One or more quadrants with prominent IRMA
R4 - PDR	Either one or all of the following: 1. Existence of neovascularization 2. Vitreous/pre-retinal hemorrhages

5.1 Binary DR Classification

This subsection reviews studies that address binary DR classification, where retinal fundus images are categorized into two classes, such as DR *versus* non-DR, healthy *versus* unhealthy, referable *versus* non-referable DR, NPDR *versus* PDR or gradable *versus* non-gradable images. Binary classification remains widely adopted due to its effectiveness in large-scale screening and referral decision support. Table 2 summarizes representative binary DR classification studies, highlighting their pre-processing techniques, datasets, methodologies, strengths and limitations.

Hu et al. [27] proposed graph adversarial transfer learning (GATL) for DR classification to address limited annotations and domain variability across retinal imaging devices. GATL integrates self-supervised learning, graph neural network (GNN)-based feature extraction and adversarial domain adaptation for intra- and inter-domain alignment. Evaluated on the APTOS 2019 and EyePACS datasets, it achieves accuracies of 94.3% and 92.7%, respectively, with high sensitivity and specificity. Its strengths include effective domain adaptation and the utilization of unlabelled data to improve classification performance. However, the method is limited by GNN assumptions, instability in adversarial training and the complexity of retinal-image characteristics. The hierarchical DL framework (HDeep) [28] is proposed for DR classification using a combination of DenseNet-121 and EfficientNet-B7 models trained on the Kaggle APTOS dataset (5,590 retinal images). The methodology reformulates the multi-class problem into four binary classification tasks and employs image enhancement and data augmentation to address class imbalance. The model achieves a Kappa score of 0.911 and an accuracy of 85.3%, outperforming comparable approaches. Its strengths include a scalable hierarchical architecture and an effective augmentation strategy, while limitations stem from severe class imbalance and the difficulty of distinguishing between similar DR severity levels.

A hybrid CNN-based DR classification model is proposed in [4] to address the challenges of manually detecting DR in fundus images. The method combines InceptionV3 and ResNet50 for feature extraction, then concatenates the features and performs classification with a CNN. The study uses 44,119 retinal images from a public dataset across five DR severity classes, applying pre-processing and data augmentation to enhance image quality and model robustness. The proposed approach achieves 96.85% accuracy and outperforms existing methods in sensitivity, specificity, precision and F1-score. Its main

strength lies in the effective integration of multiple DL architectures for improved classification performance. However, the model is evaluated on a single dataset, which may limit its generalizability to diverse clinical populations. A faster RCNN-based DR detection framework is proposed in [2] for early diagnosis by integrating fused features extracted from fundus images. The method combines histograms of oriented gradient (HOG), Shearlet transform and RCNN features into a unified feature vector after image pre-processing and noise reduction using an adaptive filter. Experiments conducted on a dataset of 21,171 fundus images achieved 98.58% accuracy, 97.12% specificity and 95.72% sensitivity. The primary strength of the approach lies in its feature-fusion strategy, which enhances detection performance compared with existing methods. However, the model still misclassifies some DR cases and requires further training and fine-tuning to improve accuracy and reduce computational cost.

A novel retinal image-quality assessment (RIQA) standard [29] is proposed that classifies retinal images into four quality grades by integrating image quality, retinal structure and DR lesion features. The method employs an attention-dilated dense network (ADD-Net), a lightweight DenseNet-based architecture that incorporates stacked dilated convolutions and attention to enhance feature representation while maintaining low computational complexity. Using the re-labeled Kaggle retinal image quality dataset, ADD-Net achieves 97.03% accuracy, 97.06% precision and 98.31% sensitivity, outperforming existing approaches. The main strengths of the approach include the establishment of a comprehensive RIQA framework and the suitability of ADD-Net for deployment on portable devices. However, the dataset exhibits class imbalance, which may introduce bias and affect the model's generalizability. Pavithra et al. [30] proposed an ensemble DL framework for DR detection using VGG19, Inception V3 and ResNet50 for feature extraction, followed by CNN-based classification. The model is evaluated on a fundus-image dataset comprising 2,076 training, 531 validation and 231 test images, with data augmentation and image enhancement applied during pre-processing. The ensemble approach achieves an accuracy of 98.47%, outperforming the individual models. The study's main strength lies in its effective integration of multiple DL architectures for binary DR classification. However, its performance depends on the availability of high-quality fundus images, which may limit generalizability in real-world clinical settings.

Table 2. Comparison of AI-based methods for binary DR classification.

Ref.	Pre-processing	Datasets	Methods	Strengths	Limitations
[27]	Resizing (512×512)	EyePACS, APTOS	GATL	Domain adaptation and graph learning improve robustness	GNN assumptions, adversarial instability, complex retinal images
[28]	CLAHE, vessel enhancement	APTOS	HDeep (DenseNet121 + EfficientNet-B7)	Hierarchical classification strategy	Sensitive to class imbalance
[4]	Histogram equalization	OCT/Fundus	Hybrid CNN (ResNet50 + InceptionV3)	Effective feature fusion	Single-dataset evaluation
[2]	Filtering, cropping	DIARETDB1, DDR, Kaggle	Faster-RCNN + HOG + Shearlet	Excellent detection accuracy	Misclassification remains for some DR cases
[29]	Cropping, resizing	DR1, DRIMDB, Kaggle	ADD-Net	Lightweight and deployable	Dataset-quality imbalance
[30]	Normalization, cropping	Kaggle	Ensemble CNN	Ensemble improves robustness	Requires high-quality images

5.2 Multi-class DR Classification

Multi-class classification in DR involves categorizing retinal images into more than two classes, such as different disease stages or severity levels. Also, Grading in DR refers to assigning a severity level or stage to the disease based on specific criteria, such as lesions, hemorrhages, exudates or microaneurysms [20]. Table 3 summarizes the key studies reviewed in this sub-section, highlighting their pre-processing techniques, datasets, methodologies, strengths and limitations in multi-class DR classification and grading.

A novel self-supervised pre-training method, MIE-DR [31], is proposed to leverage unlabelled multi-modal ophthalmic data to learn both shared and modality-specific retinal representations for DR grading. The method is evaluated on the IDRiD and Messidor datasets, with cross-dataset validation on the EyePACS-Kaggle and Messidor-2 datasets. Experimental results demonstrated superior transfer

learning (TL) performance compared with ImageNet pre-training and other multi-modal approaches, resulting in improved DR classification and grading accuracy. The main strength of MIE-DR lies in its ability to effectively exploit multi-modal information during pre-training, thereby enhancing feature representation and generalization. However, its performance depends on the availability of multi-modal data and fine-grained DR grading remains challenging. A hybrid retinal image-enhancement framework for DR detection is proposed in [19] to address low contrast, noise and uneven illumination in fundus images. The method employs image cropping and Gaussian blurring as preprocessing steps, followed by a DL-based classification model using a modified ResNet50 architecture. Evaluation on the EyePACS and MESSIDOR datasets demonstrates substantial performance improvements, with classification accuracy increasing from 73% to 92% and from 79.2% to 93.6%, respectively. The main strength of the approach lies in its effective image-enhancement strategy, which improves feature extraction and classification performance. However, the study is limited by the need for further investigation into enhancement techniques to improve contrast and image sharpness, thereby supporting more reliable DR diagnosis. In [21], the authors propose a multi-stage patch-based deep CNN (MPDCNN) that combines the InceptionV3 and Xception models for automated DR grading. The method analyzes overlapping image patches extracted from 2,290 high-resolution fundus images, sourced from hospital and public datasets, to enhance feature representation and classification accuracy. Using five-fold cross-validation, MPDCNN achieves an accuracy of 96.2%, outperforming several existing approaches. The primary strength of the model lies in its effective integration of shallow and dense CNN features, which improves grading performance. However, its ability to further differentiate high-risk PDR cases remains limited, indicating the need for more advanced DL techniques for fine-grained severity classification.

Bogacsovics et al. [32] proposed a DR detection framework that combines hand-crafted features with neural network–extracted features to improve automated fundus-image classification. Using a retinal-image dataset, the method applies pre-processing techniques, including contrast-limited adaptive histogram equalization and illumination equalization, to enhance lesion-specific feature extraction. The results demonstrate superior predictive performance compared with conventional approaches that rely solely on neural-network features. The main strength of the study lies in the effective integration of hand-crafted and deep features, which improves classification accuracy. However, the approach is limited by its reliance on a single hidden-layer architecture and requires further investigation of more advanced architectures to optimize feature fusion and overall performance. A ViT-based DL framework is proposed in [16] for automated DR severity detection by capturing long-range image dependencies through TL. The model is trained on the fine-grained annotated DR (FGADR) dataset using image normalization, data splitting, balancing and augmentation techniques, including flipping and transposition. Experimental results demonstrate superior performance compared with state-of-the-art CNN models, achieving an F1-score of 0.825. The study's strengths include the introduction of a novel ViT architecture and the investigation of data-balancing strategies. However, its effectiveness largely depends on the availability of large, annotated datasets, which may limit its applicability in data-constrained settings. She et al. [18] proposed a generative adversarial network (GAN) model to translate color fundus (CF) images into venous and late-phase fundus fluorescein angiography (FFA) images for enhanced DR screening. Pre-processing includes pixel-level image matching and filtering out poorly registered pairs, while augmentation involves random horizontal flips and rotations during training. The model is trained on 51,370 CF–venous FFA pairs and 14,644 CF–late FFA pairs, collected from 4,438 patients and is externally evaluated on the EyePACS and MESSIDOR2 datasets. Results demonstrate improved classification performance, with area under the receiver operating characteristic curve (AUC) increasing from 0.912 to 0.939 on EyePACS and from 0.952 to 0.972 on MESSIDOR2, while diabetic macular edema (DME) detection AUC improves from 0.927 to 0.974. The framework enhances DR screening by providing FFA-like information when conventional FFA is unavailable. However, the generated images may be less accurate in detecting subtle lesions and further prospective clinical validation is required.

Mahesh Patil et al. [15] proposed a DL-based DR detection framework using a ResNet-50 architecture, trained on the APTOS dataset and evaluated on the EyePACS dataset. The method incorporates image pre-processing, data augmentation and test-time augmentation (TTA) to improve classification performance and address data imbalance. Experimental results achieve a multi-class accuracy of 97.87% and a quadratic weighted kappa (QWK) score of 0.985, demonstrating strong generalization across datasets. The main strengths of the approach include its high classification accuracy, effective use of

augmentation strategies and cross-dataset evaluation capability. However, its reliance on specific benchmark datasets may limit generalizability to other DR image distributions and real-world clinical settings. A blended DL framework for DR grading is proposed in [11], which combines predictions from multiple DL models and employs 5-fold cross-validation to improve robustness and reduce overfitting. The model is trained on 33,310 fundus images from the EyePACS, APTOS, Messidor-2 and IDRiD datasets, with contrast-limited adaptive histogram equalization (CLAHE) and data augmentation applied to enhance contrast and balance the class distribution. The blended model achieves superior performance compared with individual DL architectures, demonstrating improved predictive accuracy and reliability. Its main strength lies in reducing classification bias through ensemble learning. However, its effectiveness remains dependent on dataset quality and further refinement is required to address ungradable images and misclassified samples. Athira and Nair [33] proposed a DL-based framework for automated DR detection and classification into No DR, NPDR and PDR stages using a sub-set of 3,386 fundus images from the APTOS dataset. The methodology incorporates image resizing, grayscale conversion, Gaussian blurring, circular cropping and optimized parameter tuning, followed by classification using a ResNet50-based network. The model achieves 94.7% classification accuracy and 99.8% detection accuracy, outperforming several comparative approaches. The main strengths of the framework include effective pre-processing, optimized training and enhanced detection of subtle retinal features. However, its performance may be affected by variations in image quality across the dataset and by challenges in distinguishing visually overlapping DR classes.

A DL multi-label feature extraction and classification (ML-FEC) model [34] employs pre-trained CNN architectures, including ResNet50, ResNet152 and SqueezeNet1, for multi-label lesion detection and classification in color fundus photographs (CFPs) to support early DR diagnosis. The methodology incorporates principal component analysis (PCA)-based feature extraction and data augmentation through image rotation to improve generalization. Using a fundus-image dataset, the model achieves accuracies of 94.40%, 93.67% and 91.94% for ResNet152, ResNet50 and SqueezeNet1, respectively, with low hamming loss values, demonstrating its effectiveness for clinical-screening applications. A key strength of the proposed framework is its fully automated multi-label DR detection capability, which supports scalable screening applications. However, its effectiveness is limited by the scarcity of large publicly available datasets with comprehensive lesion annotations, emphasizing the need for validation on larger, more diverse datasets to improve generalizability and clinical reliability. A graph CNN (GCNN) integrated with a variational autoencoder (VAE) framework is proposed in [20] for DR severity classification by combining local appearance features with global retinal vessel topology. The method utilizes contrast enhancement and data augmentation to improve image quality and address class imbalance and is evaluated on the Kaggle and EyePACS datasets. Experimental results achieve 90.34% accuracy on EyePACS, outperforming existing approaches by 2.85% while maintaining high sensitivity and specificity. The main strength of the approach lies in the effective integration of GCNN and VAE for enhanced feature representation. However, its adaptability to real-world clinical settings remains limited and further optimization is required to improve deployment feasibility. Mahmood et al. [35] presented a hybrid DR classification framework that integrates morphological image processing with the Inception v3 DL model to classify fundus images into five DR stages: normal, mild NPDR, moderate NPDR, severe NPDR and PDR. The approach is evaluated on the Messidor and Kaggle datasets and incorporates pre-processing steps, including blood-vessel segmentation, optic disc and macula localization and microaneurysm (MA) and hemorrhage (HEM) detection. The model achieves an overall accuracy of 96.83% with high sensitivity and specificity, demonstrating the effectiveness of combining image processing and DL techniques for automated DR diagnosis. A key strength of the framework is its ability to perform DR classification without expert supervision, while its main limitation is its dependence on image quality, which may affect diagnostic performance.

Raiaan et al. [7] proposed RetNet-10, a lightweight DL model based on a shallow CNN architecture for five-class DR classification. The model is trained on a combined dataset of 5,819 fundus images from APTOS, Messidor2 and IDRiD, with pre-processing techniques for noise and artifact removal and data augmentation using geometric, photometric and elastic deformation methods. RetNet-10 achieves a testing accuracy of 98.65%, outperforming several state-of-the-art models while maintaining low computational complexity. The main strength of the model lies in its ability to balance high classification performance with efficient processing. However, its performance may be affected by dataset imbalance and variations in image quality across the merged datasets. Mishra et al. [36] proposed a cascaded voting

ensemble deep neural network (DNN) for DR-severity classification. The method utilizes multi-level Otsu thresholding for fundus-image segmentation and combines the gray-level co-occurrence matrix (GLCM) feature extraction with dynamic flamingo optimization. Evaluated on fundus-image datasets, the model achieves high sensitivity, specificity, accuracy, precision and F1-score, demonstrating effective DR grading. Its main strength is the integration of multiple learning strategies within a robust classification framework aligned with standard grading systems, while its limitation is the increased computational complexity of the ensemble-based architecture. The fundus gradability-recognition network (FGR-Net) is proposed in [37] for automated assessment of retinal-image quality in computer-aided design (CAD) systems. The framework combines a deep autoencoder for self-supervised feature extraction and image reconstruction with a classifier to distinguish gradable from ungradable fundus images. Evaluated on retinal-image datasets, FGR-Net achieves over 89% accuracy and 87% F1-score while providing interpretable visual feedback for ophthalmologists. Its main strength lies in enhancing the interpretability of fundus-image quality assessment; however, its performance in real-time clinical settings and across diverse datasets requires further investigation.

Nazir et al. [38] proposed a weighted-average ensemble model that combines Inception-v3, VGG16 and a custom CNN for early DR detection, using the APTOS 2019 dataset (3662 training and 1992 test images). The approach incorporates image resizing and data-augmentation techniques, including rotation, zooming and flipping, to improve data diversity and reduce overfitting. The model achieves an accuracy of 95.06%, a precision of 87.88%, a recall of 83.78% and an F1-score of 85.69%. Its primary strength lies in integrating multiple DL models to develop an effective automated DR-screening system that can support clinical decision-making. However, its evaluation is limited by the use of a competition dataset without test labels, thereby restricting comprehensive performance validation. Xu et al. [25] proposed RT2Net, a multi-view joint learning framework for DR diagnosis that integrates global fundus-image features with local vascular features to improve classification performance. The methodology employs CLAHE and vascular segmentation, followed by feature extraction through dual-branch networks. Evaluated on the EyePACS and APTOS 2019 datasets, RT2Net achieves accuracies of 88.2% and 85.4%, respectively, with corresponding AUC values of 0.98 and 0.96. The model's main strength lies in its ability to leverage complementary global and local information to enhance diagnostic accuracy. However, its reliance on both feature representations may still result in misclassification in complex cases and increase model complexity. Liu [39] proposed a cascaded hybrid attention network (CHAN), which employs residual hybrid attention modules (RHAM) to extract features by combining channel attention and multi-head self-attention. The model is tested on the MESSIDOR dataset, which is partitioned into training, validation and test sets, with data augmentation techniques applied to expand the dataset to approximately 12,000 images. The training regimen includes 100 epochs with an Adam optimizer, achieving an accuracy rate of 96.34%, placing the method second among existing DR classification techniques. The main contributions include introducing the CHAN framework and effectively integrating attention mechanisms to enhance feature extraction. Limitations include the model's reliance on dataset quality and the potential challenges of optimizing attention mechanisms.

An attention-based DL framework for automated DR grading is proposed in [40], employing separate attention mechanisms for dark and bright retinal lesions using image-level labels only. The model is evaluated on the Kaggle DR dataset (35,126 training and 53,576 testing images) and achieves 83.7% accuracy and a QWK of 0.78, outperforming several existing methods. Its primary strength lies in generating interpretable attention maps that enhance clinical explainability. However, the image-decomposition process is computationally intensive and performance remains limited for certain severity classes, particularly classes 1 and 4. The DR Grading CNN (DRGCNN) [41] addresses DR grading challenges associated with class imbalance and the neglect of binocular relationships in fundus images. The model employs a CAM-EfficientNetV2-M encoder for feature extraction and a binocular feature-fusion network (BFFN) to integrate information from both eyes. Evaluation on the EyePACS (88,702 images) and Messidor-2 (1,748 images) datasets yields kappa scores of 86.62% and 86.16%, respectively, while achieving 89.06% when trained on EyePACS and tested on Messidor-2. The model enhances grading performance through balanced channel allocation and binocular feature fusion and improves interpretability using Grad-CAM. However, its performance remains affected by class imbalance and the quality of input fundus images. Naz et al. [42] proposed a robust fuzzy local information K-means clustering algorithm for DR detection that enhances conventional K-means by incorporating localized information into the clustering process. The method is evaluated on the IDRiD, Kaggle and Shiva Netralaya fundus-image datasets, with pre-processing and data-augmentation

techniques applied. Experimental results achieve 94.4% accuracy with an average execution time of 17.11 seconds and 97% accuracy under noisy conditions. The approach improves unsupervised DR detection and robustness to noise; however, its scalability to large datasets and dependence on random centroid initialization may affect efficiency and convergence.

Table 3. Comparative analysis of multi-class DR classification methods.

Ref.	Pre-processing	Dataset	Methods	Strengths	Limitations
[31]	Resizing, normalization	EyePACS, Messidor, IDRiD	Self-supervised MIE	Multi-modal representation learning	Requires multi-modal data
[19]	Cropping, Gaussian blur	EyePACS, MESSIDOR	Modified ResNet50	Image enhancement improves grading	Sensitive to image quality
[21]	Patch generation	LFH, DIARETDB, STARE, Kaggle	MPDCNN	Multi-stage feature extraction	High computational cost
[32]	CLAHE	Kaggle, Messidor, IDRiD	CNN + handcrafted features	Combines clinical and DL features	Limited architecture exploration
[16]	Cropping, normalization	FGADR	ViT	Captures global dependencies	Requires large training data
[18]	Registration and filtering	Private, EyePACS, MESSIDOR2	GAN + CNN	Novel multi-modal image translation	Clinical validation needed
[15]	Gaussian filtering	APTOS, EyePACS	ResNet50 + TTA	Excellent cross-dataset performance	Dataset dependency
[11]	CLAHE	EyePACS, APTOS, IDRiD	Blended DL	Reduces classification bias	Moderate grading performance
[33]	Resizing, grayscale, blur	APTOS	ResNet50 Hybrid	Optimized pre-processing pipeline	Image-quality variability
[34]	Rescaling, CLAHE	Retinal benchmark dataset	PCA + CNN	Multi-label lesion analysis	Limited dataset size
[20]	Contrast enhancement	EyePACS	GCNN + VAE	Exploits vascular topology	Clinical adaptation needed
[35]	CLAHE, vessel extraction	DRIVE, Messidor, Kaggle	Morphology + InceptionV3	Lesion-aware classification	Dependent on image quality
[7]	Noise removal	APTOS, Messidor2, IDRiD	RetNet-10	Lightweight architecture	Dataset imbalance
[36]	CLAHE, Gaussian thresholding	Kaggle, IDRiD, DIARETDB	Ensemble DNN	Highest grading performance	Computational complexity
[37]	None reported	EyePACS, EyeQ	FGR-Net	Image-quality assessment	Limited clinical validation
[38]	Standardization, cropping	APTOS	Ensemble CNN	Effective ensemble learning	Competition dataset limitations
[25]	CLAHE, vessel extraction	EyePACS, APTOS	RT2Net	Multi-view learning	Complexity of dual branches
[39]	Cropping, resizing	MESSIDOR	CHAN	Hybrid attention learning	Attention-optimization challenges
[40]	Cropping, normalization	EyePACS	Attention-based Xception	Improved explainability	Lower grading accuracy
[41]	Not reported	EyePACS, Messidor-2	DRGCNN	Binocular feature fusion	Class imbalance
[42]	Noise reduction	IDRiD, Kaggle	Fuzzy K-means	Robust to noisy images	Scalability concerns
[43]	Resizing	EyePACS, APTOS	Inception-ResNet Hybrid	Dual-path feature extraction	Large model complexity
[44]	Resizing, normalization	APTOS, MESSIDOR-2	DenseNet121 + GNN	Incorporates clinical metadata	Requires patient information
[45]	ZCA whitening	Kaggle	QCNET	Quantum-inspired learning	Experimental stage

Henge et al. [43] developed a multi-decision Inception-ResNet-based hybrid model for five-class DR classification using sequential and non-sequential fundus images. The architecture employs a dual-path TL framework with 172 weighted layers, processing both color and grayscale images to enhance feature representation. Evaluated on the EyePACS and APTOS datasets, the model achieves 98.1% accuracy and a 98.12% F1-score. The integration of dual-image feature mapping and chi-square refinement improves classification performance and robustness; however, the complex architecture may increase computational cost and limit deployment in resource-constrained environments. Zedadra et al. [44] proposed DRdiag, a multi-modal framework that combines a DenseNet121-based CNN with a two-layer GNN to integrate retinal-image features and patient-specific clinical information, particularly diabetes duration. Evaluated on the MESSIDOR-2 and APTOS2019 datasets, the model achieves classification accuracies of 97.6% and 98.0%, respectively. The framework demonstrates the effectiveness of graph-aware multi-modal learning for improving DR diagnosis by leveraging both visual and clinical data. However, its reliance on GNN-based modeling and patient metadata may increase system complexity and limit applicability when clinical information is unavailable. Ikram and Imran [45] proposed a hybrid quantum-classical neural network (QCNET) that combines DenseNet201 feature extraction with a 4-qubit variational quantum circuit for DR severity grading. Evaluated on the Kaggle DR dataset using

random oversampling and ZCA whitening, the model achieves 91.0% accuracy. The integration of gradient-weighted class-activation mapping (Grad-CAM) enhances interpretability by highlighting lesion-relevant retinal regions. While QCNET demonstrates the potential of quantum-enhanced learning and explainable AI, its practical applicability remains constrained by the early-stage development and scalability challenges of quantum computing technologies.

5.3 Binary and Multi-class DR Classification

This subsection discusses the literature on binary and multi-classification problems. Most research uses binary classification to distinguish between normal and DR cases, then classifies DR cases into 4 or 5 grades (Normal, Mild, Moderate, Severe and PDR). The studies reviewed in this sub-section are systematically summarized in Table 4, which provides a comparative overview of pre-processing techniques, datasets, methodologies, strengths and limitations reported in the literature.

Jaskari et al. [46] investigated uncertainty-aware DR classification using approximate Bayesian neural networks (BNNs). The study evaluates nine BNN models on EyePACS, KSSHP, Messidor-2 and APTOS datasets under a 5-class grading scheme and introduces a novel uncertainty measure linked to classifier risk. The results demonstrate that BNNs effectively estimate prediction uncertainty, while the proposed QWK-Risk metric improves performance on benchmark datasets. The study's main strength is the incorporation of uncertainty quantification into DR classification, whereas its primary limitation is reduced generalizability from benchmark datasets to real-world clinical data. Alahmadi [47] proposed a DR classification framework that integrates style and content recalibration with texture attention and spatial-normalization modules to enhance feature representation. The model is evaluated on the APTOS Kaggle dataset (3,662 retinal images) and achieves an accuracy of 0.85, demonstrating improved discrimination between healthy and DR samples. The main strength of the approach is the incorporation of attention mechanisms to enhance the learning of texture and semantic features. However, the model does not explicitly capture local contextual dependencies, which may limit its ability to model complex retinal structures. Islam et al. [48] proposed a supervised contrastive learning (SCL)-based two-stage framework that combines supervised contrastive loss for feature representation and cross-entropy loss for classification. Using the APTOS 2019 and Messidor-2 datasets with CLAHE-based pre-processing and data augmentation, the model achieves 98.36% accuracy and 98.50% AUC for binary classification and 84.36% accuracy for five-stage DR grading. The approach improves robustness and outperforms conventional CNN-based methods; however, its performance depends on effective augmentation and representation learning strategies.

A novel few-shot learning (FSL) framework called DRNet is introduced [22] using a prototypical network with episodic learning for DR detection and grading. Developed on the APTOS2019 dataset, the model achieves 99.73% accuracy, 99.82% sensitivity and 99.63% specificity for DR detection and 98.18% accuracy, 97.41% sensitivity and 99.55% specificity for DR grading. The framework demonstrates superior performance with limited training data; however, its effectiveness decreases as the number of training episodes is reduced. Tian et al. [49] proposed the fine-grained attention and knowledge-based collaborative network (FA+KC-Net) for DR grading. The method combines FA-Net to capture subtle lesion features and KC-Net to incorporate lesion-related medical knowledge. Evaluated on DDR, Messidor, APTOS and EyePACS datasets, FA+KC-Net achieves superior performance on DDR, Messidor and APTOS, demonstrating high accuracy and stability. Its main strength lies in integrating fine-grained visual features with domain knowledge, while a limitation is KC-Net's relatively lower performance, particularly in identifying DR 4 cases. Wong et al. [8] proposed a TL-based DR grading framework that combines pre-trained ShuffleNet and ResNet-18 models for feature extraction with an error correction output code (ECOC) ensemble classifier. Using the APTOS dataset, along with a combination of the EyePACS and Messidor-2 datasets and CLAHE pre-processing, the method achieves accuracies of 82% for 5-class grading and 96% for binary classification on the APTOS dataset. The approach improves performance through adaptive differential evolution (ADE)-based optimization of feature weights and network parameters. However, its effectiveness depends on image quality and may be susceptible to overfitting.

Farag et al. [3] proposed a DL-based DR severity-grading framework that uses a modified DenseNet169 with a convolutional block attention module (CBAM) and is trained on the APTOS dataset (~13,000 fundus images). Following Gaussian-filter pre-processing, the model achieves 97% accuracy in binary classification and strong multi-class grading performance, particularly for minority classes. The

approach offers efficient DR severity assessment with reduced computational complexity; however, its performance is still affected by dataset imbalance, motivating future investigation into alternative CBAM configurations and imbalanced learning strategies. Mohammadi and Nguyen [50] evaluated ChatGPT and Vertex AI for automated DR diagnosis using automated ML (AutoML) on the Messidor-2 dataset (1,740 fundus images). Image pre-processing is performed using CLAHE assisted by ChatGPT. The proposed models achieve an area under the precision-recall curve (AUPRC) of 0.81 for DR severity detection and, for DR presence detection, precision, recall and AUPRC of 84.48%, 84.48% and 0.90, respectively. The study demonstrates the potential of AI-assisted tools to support image pre-processing and analysis without coding expertise. However, its limitations include reliance on a single dataset and uncertain generalizability across diverse populations. Phridviraj et al. [51] proposed a bi-directional long short-term memory (Bi-LSTM) model combined with the multi-scale retinex with chromaticity preservation (MSRCP) pre-processing to improve DR detection from low-quality retinal-fundus images. The Bi-LSTM-MSRCP method is evaluated on the IDRiD, MESSIDOR and Kaggle datasets, achieving an accuracy of 92.35% and an F-score of 90.56%. Its primary strength is the integration of image enhancement and DL-based classification to improve detection performance, while its main limitation is its reliance on optimal tuning of MSRCP parameters.

Ashraf and Alghamdi [17] proposed a hierarchical feature-fusion network (HFF-Net), a hybrid CNN-based framework for DR screening and grading that employs hierarchical feature fusion to extract multi-scale retinal features. The model utilizes Swish activation, sparse categorical cross-entropy loss, CLAHE-based image enhancement, region-of-interest (ROI) extraction and data augmentation. Evaluated on fundus-image datasets, HFF-Net achieves 98.70% screening accuracy and 73.77% grading accuracy, outperforming competing methods. Its main strength lies in its effective multi-scale feature representation for screening and grading; however, the relatively low grading accuracy indicates challenges in distinguishing among DR severity levels. Yang et al. [52] proposed a transformer-based multiple-instance learning (TMIL) framework for DR classification, using ViT and a global instance computing block (GICB) to process high-resolution retinal images. The model is evaluated on the APTOS dataset (3,662 labeled images) for five-class grading and on the Messidor-1 dataset for binary classification. TMIL improves classification accuracy and reduces inference time by 62% compared with direct processing of high-resolution ViT. Its main strengths include efficient handling of high-resolution images and enhanced sensitivity for certain DR grades. However, the model still exhibits a false-negative rate exceeding 30%, indicating limited sensitivity in detecting early-stage DR.

Table 4. Comparative analysis of binary and multi-class DR detection methods.

Ref.	Pre-processing	Dataset	Methods	Strengths	Limitations
[46]	Resizing, variability reduction	EyePACS, KSSHP, Messidor-2, APTOS	BNNs, VGG16	Uncertainty estimation	Clinical-transferability issues
[47]	Cropping, normalization	APTOS	Texture Attention Network	Attention-guided learning	Limited contextual modeling
[48]	CLAHE	APTOS, Messidor-2	SCL + Xception	Contrastive representation learning	More complex training
[22]	Resizing	APTOS2019	DRNet (Few-Shot Learning)	Excellent performance with limited data	Sensitive to training episodes
[49]	Contrast enhancement	DDR, Messidor, APTOS, EyePACS	FA+KC-Net	Integrates medical knowledge	Reduced performance in severe grades
[8]	CLAHE	APTOS, EyePACS, Messidor-2	ECOC Ensemble	Ensemble optimization	Potential overfitting
[3]	Cropping, Gaussian filtering	APTOS2019	DenseNet169 + CBAM	Attention-enhanced classification	Dataset imbalance
[50]	CLAHE	Messidor-2	ChatGPT + Vertex AI AutoML	Low-code clinical workflow	Single-dataset evaluation
[51]	MSRCP enhancement	IDRiD, MESSIDOR, Kaggle	Bi-LSTM	Handles low-quality images	Requires parameter tuning
[17]	CLAHE, ROI extraction	IDRiD, APTOS2019	HFF-Net	Effective screening and grading	Lower grading accuracy
[52]	Cropping, patch creation	APTOS, Messidor-1	TMIL + ViT	Efficient high-resolution analysis	High false-negative rate

5.4 Lesion-based DR Classification

Some studies focus on classifying lesions in fundus images. Lesion-based classification in DR involves

categorizing retinal images by specific lesion features, such as microaneurysms, hemorrhagic spots and exudates. Table 5 presents a comprehensive summary of representative lesion-based classification studies, highlighting their pre-processing techniques, datasets, methodologies, strengths and limitations.

Lam et al. [53] proposed a DL-based retinal lesion-detection framework using CNNs trained on 1,324 ophthalmologist-labeled image patches from Kaggle retinal images. Using a sliding-window approach, the model generates lesion-probability maps and achieves AUCs of 0.94 for microaneurysms and 0.95 for exudates. The method demonstrates effective lesion detection with limited training data and supports automated retinal screening. However, it suffers from a relatively high false-positive rate due to training-data bias and requires considerable computational time for image processing. Khojasteh et al. [54] proposed a CNN-based framework for simultaneous detection of exudates, hemorrhages and microaneurysms using probability maps generated from the softmax layer. The method is evaluated on the DIARETDB1 and e-Ophtha datasets (284 fundus images) with pre-processing and data-augmentation techniques. The model achieves sensitivities of 0.96, 0.84 and 0.85 for exudates, hemorrhages and microaneurysms, respectively and accuracies of 97.3% and 86.6% on DIARETDB1 and e-Ophtha. The approach demonstrates robust DR lesion detection performance; however, it struggles to distinguish overlapping hemorrhages and microaneurysms and is affected by dataset imbalance.

Tang et al. [1] proposed a TL-based framework for neovascularization detection in PDR using AlexNet, GoogLeNet, ResNet18, ResNet50 and a hybrid ResNet18–GoogLeNet model. The method is evaluated on 1,174 retinal-image patches that have been enhanced with CLAHE. The proposed network achieves 91.57% accuracy, 85.69% sensitivity, 97.44% specificity and 97.10% precision, outperforming individual CNN models. Its main strength lies in the effective use of TL for detecting early neovascularization; however, it does not support DR severity grading and depends on the quality of ground-truth annotations. A novel DL-based method, termed HemNet [55], is proposed for automated hemorrhage detection using DIARETDB1 (89 images) and DIARETDB0 (130 images). The method employs gamma correction and a Gaussian matched filter for pre-processing and hemorrhage localization. HemNet achieves 94.21% sensitivity and 94.76% specificity, demonstrating competitive performance while reducing training time compared with deeper networks, such as AlexNet and VGG-16. Its main strength lies in its efficiency and high detection accuracy; however, its reliance on seed-point extraction may lead to missed hemorrhages when seed points are not accurately identified. A DL-based TL framework for DR detection is proposed in [56] using a modified VGG16 model for feature extraction and a logistic regression classifier. The approach is evaluated on an Indian DR dataset containing four lesion classes: microaneurysms, soft exudates, hard exudates and hemorrhages. Image pre-processing and data-augmentation techniques are employed to improve performance, achieving an accuracy of 90.4%. The main strength of the method is its reliable screening capability through CNN-based feature learning, while its limitations include insufficient noise reduction, lack of metadata integration and limited feature-selection strategies.

Table 5. Comparison of lesion-based approaches for automated DR detection.

Ref.	Pre-processing	Dataset	Methods	Strengths	Limitations
[53]	Normalization	Kaggle, E-Ophtha	CNN	Excellent lesion detection	Computationally intensive
[54]	Contrast enhancement	DIARETDB1, E-Ophtha	CNN	Multi-lesion detection	Difficulty with overlapping lesions
[1]	CLAHE	Messidor, DIARETDB0	TL	Detects neovascularization	No grading capability
[55]	Gamma correction	DIARETDB0/1	HemNet	Lightweight model	Seed-point dependency
[56]	Scaling, color correction	IDRiD	VGG16 / VGG19 TL	Strong lesion representation	Needs better feature selection

6. BENCHMARK DATASETS FOR DR DETECTION

This section summarizes key datasets widely used in DR research, particularly for classification tasks, such as severity grading and lesion identification. These datasets consist of retinal-fundus images captured under varying conditions, with differences in resolution, field of view (FOV) and annotation detail and include labels for pathologies, such as microaneurysms, hemorrhages, exudates and DME. Table 6 provides a detailed comparison of various DR datasets commonly used in research. Commonly used datasets include Messidor [57] (1,200 images with DR and DME grading), DIARETDB0/1 (130 and 89 images with expert lesion annotations) and Kaggle DR (EyePACS) [58], a large-scale dataset of 35,126 images with five-class DR grading, though affected by variations in image quality. Messidor-2

[57][59] extends Messidor with 1,748 images from 874 examinations, but lacks standardized DR annotations.

Other datasets provide more specialized or detailed annotations. E-ophtha [60] focuses on lesion-level labeling for microaneurysms and exudates, while DRiDB [61] offers detailed annotations of lesions and anatomical structures from multiple experts. IDRiD [62] includes 516 high-resolution images with both DR grading and pixel-level lesion annotations, supporting multi-task learning. APTOS 2019 [63] contains 3,662 images labeled according to ICDRSS severity levels under diverse real-world conditions. Larger and more recent datasets include DDR [64] (13,673 images from multiple hospitals, including poor-quality cases) and FGADR [65] (2,842 images with fine-grained lesion annotations and high-quality grading labels from multiple ophthalmologists), both of which enable more robust evaluation of DR classification models.

Table 6. Publicly available dataset for DR detection.

Dataset Name	**Images**	**No DR**	**Mild DR**	**Moderate DR**	**Severe DR**	**PDR**	**Fundus Camera**	**Annotation Level**	**Usage**
Messidor [57]	1200	546	153	247	254	-	Topcon TRC NW6 non-mydriatic, 45° FOV	Image-level	DR grading
DIARETDB0/ DIARETDB1 [66]	130/89	20 / 5	110/84				Zeiss FF450 + 50° FOV	Pixel-level and image-level	DR grading, Lesion Detection
Kaggle DR (EyePACS) [58]	88702 (35126 train)	73726 25810	7351 2443	7025 5292	708 873	1892 708	Various cameras (Centervue DRS, Optovue iCam, Canon CR1 / DGi / CR2, Topcon NW)	Image-level	DR grading
Messidor-2 [57]	1748	874	174	524	87	89	Topcon TRC NW6 non-mydriatic, 45° FOV	Image-level	DR grading
E-Ophtha [60]	463	-	-	-	-	-	Topcon TRC-NW6S, Canon CR-DGi	Pixel-level	DR diagnosis, Lesion Detection
DRiDB [61]	50	14	36				Zeiss VISUCAM 200, 45° FOV	Pixel-level and image-level	DR diagnosis, Lesion Detection
IDRiD [62]	516	134	20	193	65	104	Kowa VX-10 Alpha, 50° FOV	Pixel-level and image-level	DR grading, Lesion detection
APTOS 2019 [63]	5593	1805	370	999	193	295	Various fundus cameras	Image-level	DR grading
DDR [64]	13673	6266	630	4477	236	913	Various fundus cameras	Image-level	DR grading, Lesion Detection
FGADR [65]	2842	143	125	566	105	61	Not specified	Pixel-level and image-level	DR grading, Lesion Detection

Although these benchmark datasets have significantly advanced AI-based DR research, several important limitations remain regarding dataset diversity, annotation consistency and real-world generalization. Despite the availability of large benchmark datasets, significant dataset bias remains a major challenge in AI-driven DR detection. Public datasets often differ substantially in their demographic distributions, ethnicities, image acquisition devices, illumination conditions and annotation protocols. Such variability can undermine cross-dataset generalization and degrade performance when models are deployed in real-world clinical environments. Furthermore, annotation inconsistency among ophthalmologists introduces label uncertainty, particularly for intermediate DR stages. These limitations highlight the need for standardized evaluation protocols, diverse multi-center datasets and fairness-aware learning strategies.

7. DISCUSSION AND RESEARCH GAPS

The literature reviewed in this paper clearly demonstrates that AI–driven solutions for DR detection have undergone rapid methodological evolution, moving from coarse binary screening systems to sophisticated multi-task frameworks that integrate classification, lesion detection, segmentation and clinical interpretability. This progression reflects a broader shift in medical-image analysis from purely performance-driven models toward systems that aim to approximate clinical reasoning, scalability and trustworthiness.

Binary DR classification remains the most widely adopted paradigm, largely due to its direct applicability in population-level screening programs. Studies focusing on DR *versus* non-DR or referable *versus* non-referable DR consistently report high accuracy, sensitivity and specificity, as

demonstrated by GATL [27], HDeep [28] and ensemble CNN approaches [30]. These models effectively reduce ophthalmologists' workload by filtering out normal cases, thereby enabling targeted referral. However, a closer examination reveals that these impressive metrics are often achieved under controlled experimental settings with curated datasets, such as EyePACS and APTOS. While domain-adaptation strategies, such as GATL [27], attempt to mitigate inter-dataset variability, most binary classifiers still struggle when exposed to heterogeneous clinical environments characterized by diverse imaging devices, inconsistent illumination and variable patient demographics. Moreover, binary classification inherently oversimplifies DR pathology, failing to capture disease-progression dynamics that are essential for personalized treatment planning.

Multi-class DR classification and grading represent a more clinically meaningful task, as they directly correspond to treatment decisions and follow-up strategies. The reviewed studies highlight a clear trend toward deeper architectures, attention mechanisms and transformer-based models to address the subtle inter-class differences among DR stages. Self-supervised and multi-modal pre-training strategies, such as MIE-DR [31], demonstrate that leveraging unlabeled ophthalmic data significantly enhances feature representation, particularly when labeled datasets are limited. Similarly, ViT-based approaches [16] [52] successfully model long-range spatial dependencies, which are critical for capturing global retinal context. Attention-driven CNNs [40] [39] further improve discriminative power by focusing on lesion-relevant regions. Despite these advances, grading performance remains inconsistent across DR stages. Intermediate classes (mild and moderate NPDR) are frequently misclassified due to overlapping lesion characteristics and gradual disease transitions. This limitation is repeatedly observed even in high-performing systems, such as DRGCNN [41]and RT2Net [25]. The problem is exacerbated by severe class imbalance, where minority classes receive limited representation during training, leading to biased decision boundaries.

Lesion-based classification approaches address a critical limitation of pure image-level classifiers by explicitly modeling pathological features, such as microaneurysms, hemorrhages and exudates. Early CNN-based lesion detectors [53] [54] established the feasibility of automated lesion recognition with relatively small datasets, highlighting the potential for explainable DR screening systems. However, lesion-based methods often operate in isolation, focusing on individual lesion types without fully integrating their spatial distribution, co-occurrence patterns or clinical significance, while newer methods incorporate TL and hybrid architectures [56][55]. Lesion overlap and visual similarity—particularly between microaneurysms and hemorrhages—remain persistent challenges. Consequently, lesion-based classification alone is insufficient for reliable severity grading, reinforcing the need for integrated frameworks.

Another important consideration when interpreting the literature is the heterogeneity of the diagnostic tasks across studies. Although many works are collectively categorized under DR detection, they often target fundamentally different objectives, including binary DR screening, referable *versus* non-referable DR classification, multi-class severity grading, lesion detection, image-quality assessment and segmentation-assisted diagnosis. Consequently, performance metrics reported across these studies are not always directly comparable, because task complexity, annotation requirements, class distributions and evaluation protocols differ substantially. Binary screening tasks generally achieve higher accuracy due to simpler decision boundaries, whereas multi-class grading and lesion-level analysis present greater challenges related to inter-class similarity, annotation variability and clinical ambiguity. Therefore, reported performance should be interpreted in the context of the specific diagnostic objective and experimental setting, rather than as a universally comparable indicator of methodological superiority.

Despite these advances, several critical challenges persist. Most studies rely on a limited set of public datasets (EyePACS, APTOS, IDRiD, MESSIDOR), leading to weak cross-dataset generalization and performance degradation on unseen clinical data [15], [46]. Although many studies report high classification accuracy and AUC values, direct comparison among models remains difficult, because the evaluated methods use different datasets, pre-processing techniques, train-test splits, grading protocols and augmentation strategies. Consequently, reported performance metrics may not provide a fully fair or standardized comparison of state-of-the-art methods. Future research should focus on unified benchmarking frameworks and standardized evaluation protocols to enable more reliable comparative analysis. Reproducibility remains a major challenge in AI-based DR research. Many studies do not provide publicly available source codes, detailed pre-processing pipelines or standardized train-validation-test splits, making independent validation difficult. Furthermore, inconsistent

evaluation metrics and dataset protocols complicate fair comparison among models. Establishing open benchmarking frameworks and reproducible evaluation standards is essential for improving transparency, facilitating independent verification and accelerating clinical translation. Class imbalance continues to hinder reliable detection of early and intermediate DR stages, increasing the risk of false negatives. Reported high-performance metrics should be interpreted cautiously, as several studies rely on highly imbalanced datasets, limited external validation or extensive augmentation strategies that may artificially inflate performance. Furthermore, differences in experimental protocols and evaluation settings reduce reproducibility and complicate fair comparison across studies. These issues highlight the importance of robust validation using diverse multi-center datasets and standardized testing frameworks. Although explainability techniques, such as attention maps and Grad-CAM, improve interpretability, many models still lack clinically meaningful explanations aligned with ophthalmological reasoning. Furthermore, high computational complexity limits deployment in resource-constrained environments, despite the emergence of lightweight models, such as RetNet-10 and ADD-Net. Finally, most approaches rely solely on fundus images, with limited use of multi-modal and longitudinal data. Addressing these challenges is essential to bridge the gap between experimental performance and real-world clinical deployment of AI-based DR classification systems.

Although many AI-based DR detection systems achieve high performance on benchmark datasets, their translation into routine clinical practice remains limited. Most studies rely on retrospective evaluations using public datasets and lack prospective clinical validation across diverse patient populations and healthcare settings. In addition, substantial variation exists in validation protocols across studies, including differences in train-validation-test splits, cross-validation strategies, external validation practices and performance-reporting standards, making reliable comparison and clinical assessment difficult. Furthermore, regulatory approval of AI-based medical devices requires robust external validation, reliability assessment, explainability and evidence of clinical benefit. Practical deployment also faces challenges related to integration with existing healthcare workflows, interoperability with electronic health record systems, data privacy and security requirements and clinician trust in automated decision-support systems. Consequently, future research should prioritize clinically oriented validation studies, standardized regulatory-compliant evaluation protocols and human-centered AI frameworks to facilitate the safe and effective adoption of AI-driven DR screening systems in real-world clinical environments.

Recent research has also explored emerging paradigms beyond conventional CNN- and transformer-based architectures. Emerging approaches, such as multi-modal learning and quantum-inspired neural networks, demonstrate promising capability for improving DR classification performance. Multi-modal systems can integrate complementary information from fundus images, OCT scans and clinical metadata, thereby enhancing robustness and diagnostic precision. However, these methods typically require larger datasets, increased computational resources and more complex training pipelines. Similarly, quantum-inspired models remain in early experimental stages and currently face scalability and hardware limitations that restrict practical clinical deployment. Despite these challenges, such approaches represent promising directions for future research toward more comprehensive and clinically reliable DR detection systems.

8. Conclusion and Future Directions

This review examines AI-based approaches for DR classification, including binary detection and multi-class severity grading. Recent advances in deep learning, particularly CNNs, attention mechanisms and transformer-based models, have achieved high classification performance, in some cases approaching expert-level accuracy. However, key challenges remain. Many models suffer from limited generalization, as they rely heavily on a few benchmark datasets, leading to performance drops in real-world settings. Class imbalance further affects the reliable detection of early and intermediate DR stages. Additionally, most classification models rely on global image features and integrate only limited lesion-level information, while current interpretability methods often fall short of clinical expectations. From a practical perspective, computational complexity limits deployment, especially for transformer-based and ensemble models and achieving efficiency–accuracy balance remains challenging. Moreover, most approaches rely on single-modality fundus images, restricting their ability to capture disease variability. In conclusion, despite strong progress, bridging the gap between high experimental accuracy and real-

world clinical deployment remains a key challenge for DR classification systems.

Future research should focus on developing generalized, clinically reliable DR detection systems that use domain adaptation and federated learning strategies to preserve data privacy while improving cross-center generalization. Self-supervised and foundation-model approaches may reduce dependence on large, labeled datasets, while multi-modal learning can integrate fundus imaging with OCT and clinical metadata to improve diagnostic robustness. In addition, lightweight edge AI architectures are needed for deployment in resource-constrained screening environments. Standardized benchmarking frameworks and fairness-aware evaluation protocols should also be established to improve reproducibility and clinical reliability.

REFERENCES

[1] M. C. S. Tang, S. S. Teoh, H. Ibrahim and Z. Embong, "A Deep Learning Approach for the Detection of Neovascularization in Fundus Images Using Transfer Learning," IEEE Access, vol. 10, pp. 20247–20258, DOI: 10.1109/ACCESS.2022.3151644, 2022.

[2] Md. Nur-A-Alam, Md. M. K. Nasir, M. Ahsan, Md. A. Based, J. Haider and S. Palani, "A Faster RCNN-based Diabetic Retinopathy Detection Method Using Fused Features From Retina Images," IEEE Access, vol. 11, pp. 124331–124349, DOI: 10.1109/ACCESS.2023.3330104, 2023.

[3] M. M. Farag, M. Fouad and A. T. Abdel-Hamid, "Automatic Severity Classification of Diabetic Retinopathy Based on DenseNet and Convolutional Block Attention Module," IEEE Access, vol. 10, pp. 38299–38308, DOI: 10.1109/ACCESS.2022.3165193, 2022.

[4] G. Ali, A. Dastgir, M. W. Iqbal, M. Anwar and M. Faheem, "A Hybrid Convolutional Neural Network Model for Automatic Diabetic Retinopathy Classification From Fundus Images," IEEE J. of Translational Engineering in Health and Medicine, vol. 11, pp. 341–350, DOI: 10.1109/JTEHM.2023.3282104, 2023.

[5] M. Shafiq, Q. Fan, F. H. Alghamedy and W. J. Obidallah, "DualEye-FeatureNet: A Dual-stream Feature Transfer Framework for Multi-modal Ophthalmic Image Classification," IEEE Access, vol. 12, pp. 143985–144008, DOI: 10.1109/ACCESS.2024.3469244, 2024.

[6] M. N. Hasan, M. E. R. Pial, S. Das, N. Siddique and H. Wang, "DIA-VXNET: A Framework for Automated Diabetic Eye Disease Detection Using Transfer Learning with Feature Fusion Network," Biomedical Signal Processing and Control, vol. 100, p. 106907, Feb. 2025.

[7] M. A. K. Raiaan et al., "A Lightweight Robust Deep Learning Model Gained High Accuracy in Classifying a Wide Range of Diabetic Retinopathy Images," IEEE Access, vol. 11, pp. 42361–42388, DOI: 10.1109/ACCESS.2023.3272228, 2023.

[8] W. K. Wong, F. H. Juwono and C. Apriono, "Diabetic Retinopathy Detection and Grading: A Transfer Learning Approach Using Simultaneous Parameter Optimization and Feature-weighted ECOC Ensemble," IEEE Access, vol. 11, pp. 83004–83016, DOI: 10.1109/ACCESS.2023.3301618, 2023.

[9] T. Palaniswamy and M. Vellingiri, "Internet of Things and Deep Learning Enabled Diabetic Retinopathy Diagnosis Using Retinal Fundus Images," IEEE Access, vol. 11, pp. 27590–27601, 2023.

[10] M. Nawal Sad-Houari Hicham Reguieg, Chaimaa Bachiri, "Automated Diabetes Disease Prediction System Based on Risk Factors Assessment: Taking Charge of Your Health," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 09, no. 04, pp. 377–394, DOI: 10.5455/jjcit.71-1694107771, 2023.

[11] F. C. Monteiro, "Diabetic Retinopathy Grading Using Blended Deep Learning," Procedia Computer Science, vol. 219, pp. 1097–1104, DOI: 10.1016/j.procs.2023.01.389, 2022.

[12] B. N. Jagadesh, M. G. Karthik, D. Siri, S. K. K. Shareef, S. V. Mantena and R. Vatambeti, "Segmentation Using the IC2T Model and Classification of Diabetic Retinopathy Using the Rock Hyrax Swarm-based Coordination Attention Mechanism," IEEE Access, vol. 11, pp. 124441–124458, DOI: 10.1109/ACCESS.2023.3330436, 2023.

[13] S. Samira Ait Kaci Azzou, D. Boukredera and S. Baouz, "A Hybrid CNN-Transformer Approach for Precise Three-class Diabetic Retinopathy Classification," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 11, no. 03, pp. 279–299, DOI: 10.5455/jjcit.71-1738691198, 2025.

[14] Q. Abbas, M. E. A. Ibrahim and A. R. Baig, "Transfer Learning-based Computer-aided Diagnosis System for Predicting Grades of Diabetic Retinopathy," Computers, Materials & Continua, vol. 71, no. 3, pp. 4573–4590, doi: 10.32604/cmc.2022.023670, Jan. 2022.

[15] M. S. Patil, S. Chickerur, C. Abhimalya, A. Naik, N. Kumari and S. Maurya, "Effective Deep Learning Data Augmentation Techniques for Diabetic Retinopathy Classification," Proc. of the Int. Conf. on Machine Learning and Data Engineering, vol. 218, pp. 1156–1165, Jan. 2023.

[16] W. Nazih, A. O. Aseeri, O. Y. Atallah and S. El-Sappagh, "Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-based Retina Images," IEEE Access, vol. 11, pp. 117546–117561, DOI: 10.1109/ACCESS.2023.3326528, 2023.

[17] M. H. Ashraf and H. Alghamdi, "HFF-Net: A Hybrid Convolutional Neural Network for Diabetic

Retinopathy Screening and Grading," Biomedical Technology, vol. 8, pp. 50–64, Dec. 2024.

[18] D. Shi et al., "Translation of Color Fundus Photography into Fluorescein Angiography Using Deep Learning for Enhanced Diabetic Retinopathy Screening," Ophthalmology Science, vol. 3, no. 4, p. 100401, Dec. 2023.

[19] S. H. Abbood, H. N. A. Hamed, M. S. M. Rahim, A. Rehman, T. Saba and S. A. Bahaj, "Hybrid Retinal Image Enhancement Algorithm for Diabetic Retinopathy Diagnostic Using Deep Learning Model," IEEE Access, vol. 10, pp. 73079–73086, DOI: 10.1109/ACCESS.2022.3189374, 2022.

[20] S. Sundar and S. Sumathy, "Classification of Diabetic Retinopathy Disease Levels by Extracting Topological Features Using Graph Neural Networks," IEEE Access, vol. 11, pp. 51435–51444, DOI: 10.1109/ACCESS.2023.3279393, 2023.

[21] V. Deepa, C. S. Kumar and T. Cherian, "Ensemble of Multi-stage Deep Convolutional Neural Networks for Automated Grading of Diabetic Retinopathy Using Image Patches," J. of King Saud Uni. - Computer and Information Sciences, vol. 34, no. 8, Part B, pp. 6255–6265, Sep. 2022.

[22] M. Murugappan, N. B. Prakash, R. Jeya, A. Mohanarathinam, G. R. Hemalakshmi and M. Mahmud, "A Novel Few-shot Classification Framework for Diabetic Retinopathy Detection and Grading," Measurement, vol. 200, p. 111485, DOI: 10.1016/j.measurement.2022.111485, Aug. 2022.

[23] J. Amin, I. Shazadi, M. Sharif, M. Yasmin, N. A. Almujally and Y. Nam, "Localization and Grading of NPDR Lesions Using ResNet-18-YOLOv8 Model and Informative Features Selection for DR Classification Based on Transfer Learning," Heliyon, vol. 10, no. 10, p. e30954, May 2024.

[24] K. Aurangzeb, R. S. Alharthi, S. I. Haider and M. Alhussein, "Systematic Development of AI-enabled Diagnostic Systems for Glaucoma and Diabetic Retinopathy," IEEE Access, vol. 11, pp. 105069–105081, DOI: 10.1109/ACCESS.2023.3317348, 2023.

[25] X. Xu, D. Liu, G. Huang, M. Wang, M. Lei and Y. Jia, "Computer Aided Diagnosis of Diabetic Retinopathy Based on Multi-view Joint Learning," Computers in Biology and Medicine, vol. 174, p. 108428, May 2024.

[26] C. P. Wilkinson et al., "Proposed International Clinical Diabetic Retinopathy and Diabetic Macular Edema Disease Severity Scales," Ophthalmology, vol. 110, no. 9, pp. 1677–1682, Sep. 2003.

[27] J. Hu, H. Wang, L. Wang and Y. Lu, "Graph Adversarial Transfer Learning for Diabetic Retinopathy Classification," IEEE Access, vol. 10, pp. 119071–119083, DOI: 10.1109/ACCESS.2022.3220776, 2022.

[28] T. F. de Sousa and C. G. Camilo, "HDeep: Hierarchical Deep Learning Combination for Detection of Diabetic Retinopathy," Proc. of the Int. Neural Network Society Workshop on Deep Learning Innovations and Applications (INNS DLIA 2023), vol. 222, pp. 425–434, Jan. 2023.

[29] P. Zhang et al., "Image Quality Assessment of Diabetic Retinopathy Based on ADD-Net," IEEE Access, vol. 11, pp. 105130–105139, DOI: 10.1109/ACCESS.2023.3318876, 2023.

[30] S. Pavithra, D. Jaladi and K. Tamilarasi, "Optical Imaging for Diabetic Retinopathy Diagnosis and Detection Using Ensemble Models," Photodiagnosis and Photodynamic Therapy, vol. 48, p. 104259, Aug. 2024.

[31] Á. S. Hervella, J. Rouco, J. Novo and M. Ortega, "Multimodal Image Encoding Pre-training for Diabetic Retinopathy Grading," Computers in Biology and Medicine, vol. 143, p. 105302, Apr. 2022.

[32] G. Bogacsovics, J. Toth, A. Hajdu and B. Harangi, "Enhancing CNNs through the Use of Hand-crafted Features in Automated Fundus Image Classification," Biomedical Signal Processing and Control, vol. 76, p. 103685, DOI: 10.1016/j.bspc.2022.103685, Jul. 2022.

[33] T. R. Athira and J. J. Nair, "Diabetic Retinopathy Grading From Color Fundus Images: An Autotuned Deep Learning Approach," Proc. of the Int. Conf. on Machine Learning and Data Engineering, vol. 218, pp. 1055–1066, DOI: 10.1016/j.procs.2023.01.085, Jan. 2023.

[34] T. M. Usman, Y. K. Saheed, D. Ignace and A. Nsang, "Diabetic Retinopathy Detection Using Principal Component Analysis Multi-label Feature Extraction and Classification," Int. J. of Cognitive Computing in Eng., vol. 4, pp. 78–88, DOI: 10.1016/j.ijcce.2023.02.002, Jun. 2023.

[35] M. A. I. Mahmood, N. Aktar and Md. F. Kader, "A Hybrid Approach for Diagnosing Diabetic Retinopathy from Fundus Image Exploiting Deep Features," Heliyon, vol. 9, no. 9, p. e19625, Sep. 2023.

[36] A. Mishra, M. Pandey and L. Singh, "Segmentation and Classification of Diabetic Retinopathy Using Ensemble Deep Neural Network," Proc. of the 5th Int. Conf. on Innovative Data Communication Technologies and Application (ICIDCA 2024), vol. 233, pp. 226–234, DOI: 10.1016/j.procs.2024.03.212, Jan. 2024.

[37] S. Khalid, H. A. Rashwan, S. Abdulwahab, M. Abdel-Nasser, F. M. Quiroga and D. Puig, "FGR-Net: Interpretable Fundus Image Gradeability Classification Based on Deep Reconstruction Learning," Expert Systems with Applications, vol. 238, p. 121644, DOI: 10.1016/j.eswa.2023.121644, Mar. 2024.

[38] K. Nazir, J. Kim and Y.-C. Byun, "Enhancing Early-stage Diabetic Retinopathy Detection Using a Weighted Ensemble of Deep Neural Networks," IEEE Access, vol. 12, pp. 113565–113579, 2024.

[39] Y. Liu, "Automatic Grading Diabetic Retinopathy in Color Fundus Image: Cascaded Hybrid Attention Network," Journal of Radiation Research and Applied Sciences, vol. 17, no. 4, p. 101160, Dec. 2024.

[40] R. Romero-Oraá, M. Herrero-Tudela, M. I. López, R. Hornero and M. García, "Attention-based Deep Learning Framework for Automatic Fundus Image Processing to Aid in Diabetic Retinopathy Grading," Computer Methods and Programs in Biomedicine, vol. 249, p. 108160, Jun. 2024.

[41] Z. Hai et al., "A Novel Approach for Intelligent Diagnosis and Grading of Diabetic Retinopathy," Computers in Biology and Medicine, vol. 172, p. 108246, Apr. 2024.

[42] H. Naz, T. Saba, F. S. Alamri, A. S. Almasoud and A. Rehman, "An Improved Robust Fuzzy Local Information K-means Clustering Algorithm for Diabetic Retinopathy Detection," IEEE Access, vol. 12, pp. 78611–78623, DOI: 10.1109/ACCESS.2024.3392032, 2024.

[43] S. K. Henge, N. R. Viraati, M. Alhussein, A. S. Kushwaha, K. Aurangzeb and R. Singh, "Detection of Diabetic Retinopathy Using a Multi-decision Inception-ResNet-blended Hybrid Model," IEEE Access, vol. 13, pp. 8988–9005, DOI: 10.1109/ACCESS.2024.3525154, 2025.

[44] A. Zedadra, O. Zedadra, M. Yassine Salah-Salah and A. Guerrieri, "Graph-aware Multimodal Deep Learning for Classification of Diabetic Retinopathy Images," IEEE Access, vol. 13, pp. 74799–74810, 2025.

[45] A. Ikram and A. Imran, "Fundus Imaging Based Quantum Computing: Grading of Severity of Vision Threatening Diabetic Retinopathy," Informatics in Medicine Unlocked, vol. 61, p. 101744, DOI: 10.1016/j.imu.2026.101744, Mar. 2026.

[46] J. Jaskari et al., "Uncertainty-aware Deep Learning Methods for Robust Diabetic Retinopathy Classification," IEEE Access, vol. 10, pp. 76669–76681, DOI: 10.1109/ACCESS.2022.3192024, 2022.

[47] M. D. Alahmadi, "Texture Attention Network for Diabetic Retinopathy Classification," IEEE Access, vol. 10, pp. 55522–55532, DOI: 10.1109/ACCESS.2022.3177651, 2022.

[48] M. R. Islam et al., "Applying Supervised Contrastive Learning for The Detection of Diabetic Retinopathy and Its Severity Levels from Fundus Images," Computers in Biology and Medicine, vol. 146, p. 105602, DOI: 10.1016/j.compbiomed.2022.105602, Jul. 2022.

[49] M. Tian, H. Wang, Y. Sun, S. Wu, Q. Tang and M. Zhang, "Fine-grained Attention & Knowledge-based Collaborative Network for Diabetic Retinopathy Grading," Heliyon, vol. 9, no. 7, p. e17217, Jul. 2023.

[50] S. S. Mohammadi and Q. D. Nguyen, "A User-friendly Approach for the Diagnosis of Diabetic Retinopathy Using ChatGPT and Automated Machine Learning," Ophthalmology Science, vol. 4, no. 4, p. 100495, 2024.

[51] M. S. B. Phridviraj, R. Bhukya, S. Madugula, A. Manjula, S. Vodithala and M. S. Waseem, "A Bi-directional Long Short-term Memory-based Diabetic Retinopathy Detection Model Using Retinal Fundus Images," Healthcare Analysis, vol. 3, p. 100174, DOI: 10.1016/j.health.2023.100174, Nov. 2023.

[52] Y. Yang, Z. Cai, S. Qiu and P. Xu, "A Novel Transformer Model with Multiple Instance Learning for Diabetic Retinopathy Classification," IEEE Access, vol. 12, pp. 6768–6776, 2024.

[53] C. Lam, C. Yu, L. Huang and D. Rubin, "Retinal Lesion Detection with Deep Learning Using Image Patches," Investigative Ophthalmology & Visual Science, vol. 59, no. 1, p. 590, Jan. 2018.

[54] P. Khojasteh, B. Aliahmad and D. K. Kumar, "Fundus Images Analysis Using Deep Features for Detection of Exudates, Hemorrhages and Microaneurysms," BMC Ophthalmology, vol. 18, no. 1, p. 288, Dec. 2018.

[55] T. Aziz, C. Charoenlarpnopparut and S. Mahapakulchai, "Deep Learning-based Hemorrhage Detection for Diabetic Retinopathy Screening," Scientific Reports, vol. 13, no. 1, p. 1479, Jan. 2023.

[56] D. Singh and D. C. Dobhal, "A Deep Learning-based Transfer Learning Approach Fine-tuned for Detecting Diabetic Retinopathy," Proc. of the 5th Int. Conf. on Innovative Data Communication Technologies and Application (ICIDCA 2024), vol. 233, pp. 444–453, DOI: 10.1016/j.procs.2024.03.234, Jan. 2024.

[57] E. Decencière et al., "Feedback on a Publicly Distributed Image Database: The Messidor Database," Image Analysis & Stereology, vol. 33, no. 3, Art. no. 3, DOI: 10.5566/ias.1155, Aug. 2014.

[58] Emma Dugas, Jared, Jorge and Will Cukierski, "Diabetic Retinopathy Detection," Diabetic Retinopathy Detection, [Online], Available: https://kaggle.com/diabetic-retinopathy-detection, Accessed: Dec. 14, 2024.

[59] G. P. Maffre Gervais Gauthier et al., "Messidor-2," ADCIS, [Online], Available: https://www.adcis.net/en/third-party/messidor2/, Accessed: Dec. 14, 2024.

[60] E. Decencière et al., "TeleOphta: Machine Learning and Image Processing Methods for Teleophthalmology," ANR TECSAN - Technologies for Health and Autonomy, vol. 34, no. 2, pp. 196–203, DOI: 10.1016/j.irbm.2013.01.010, Apr. 2013.

[61] P. Prentašić et al., "Diabetic Retinopathy Image Database (DRiDB): A New Database for Diabetic Retinopathy Screening Programs Research," Proc. of the 2013 8th Int. Symposium on Image and Signal Processing and Analysis (ISPA), pp. 711–716, DOI: 10.1109/ISPA.2013.6703830, 2013.

[62] S. P. Prasanna Porwal, "Indian Diabetic Retinopathy Image Dataset (IDRiD)," IEEE DataPort, DOI: 10.21227/H25W98, 2018.

[63] M. Karthik and S. Dane, "APTOS 2019 Blindness Detection," [Online], Available: https://kaggle.com/aptos2019-blindness-detection, Accessed: Dec. 14, 2024.

[64] T. Li, Y. Gao, K. Wang, S. Guo, H. Liu and H. Kang, "Diagnostic Assessment of Deep Learning Algorithms for Diabetic Retinopathy Screening," Information ScienceS, vol. 501, pp. 511–522, DOI: https://doi.org/10.1016/j.ins.2019.06.011, 2019.

[65] Y. Zhou, B. Wang, L. Huang, S. Cui and L. Shao, "A Benchmark for Studying Diabetic Retinopathy: Segmentation, Grading and Transferability," IEEE Transactions on Medical Imaging, vol. 40, no. 3, pp. 818–828, DOI: 10.1109/TMI.2020.3037771, Mar. 2021.

[66] T. Kauppi et al., "DIARETDB1 Diabetic Retinopathy Database and Evaluation Protocol," [Online], Available: https://webpages.tuni.fi/vision/public_data/publications/bmvc2007_diaretdb1.pdf, 2007.

ملخص البحث:

يُعدّ اعتلال الشّبكية السُّكّري سبباً رئيسياً لضعّف البصر الّذي يمكن الوقاية منه، ممّا يستدعي توفير حلول فحص موثوقة وقابلة للتّوسّع. وقد ساهم الذّكاء الاصطناعي – ولا سيما التّعلُّم الآلي والتّعلُّم العميق – بشكلٍ كبير في أتمتة تصنيف اعتلال الشّبكية السّكّري استناداً إلى صور قاع العين.

تقدّم هذه الورقة مراجعة منهجية لأساليب تصنيف اعتلال الشّبكية السّكّري القائمة على الذّكاء الاصطناعي، مع التّركيز على التّصنيف الثّنائي، وتصنيف درجات الشّدّة متعدّد الفئات، والطّرق القائمة على الآفات. كما تطرح تصنيفاً لهذه الأساليب بناءً على أهدافها السّريرية. وتغطّي المراجعة أحدث النّماذج المتطوّرة، بما في ذلك الشّبكات العصبية التّلافيفية، ومحوّلات الرّؤية، وآليات الانتباه، وتقنيات النّماذج المجمّعة، بالاضافة إلى استراتيجيات المعالجة المسبقة وبروتوكولات التّقييم الخاصّة بها. كما يتمّ تحليل مجموعات البيانات المرجعية الشّائعة من حيث توزيع الفئات وتأثيرها على الأداء.

وعلى الرّغم من أنّ النّماذج الحالية تحقّق دقّة عالية في مجموعات البيانات القياسية، فإنّ التّحدّيات لا تزال قائمةً فيما يتعلّق بالقُدرة على التّعميم، وعدم توازن الفئات، والموثوقية في الكشْف عن المرض في مراحله المبكّرة. وتجدر الإشارة إلى أنّ الأساليب القائمة على الآفات تعزّز قابلية التّفسير، إلّا أنّها غالباً لا تكون مُدمجةً بالكامل في أطر تصنيف درجات المرض.

وبشكلٍ عام، تسلّط هذه المراجعة الضّوء على القيود الرّئيسية وتحدّد التّوجُّهات المستقبلية نحو تطوير أنظمة قوية وقابلة للتّفسير وقابلة للتّطبيق السّريري لتصنيف اعتلال الشّبكية السّكّري في سياق الفحص الميداني الفعلي.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

DUAL-LEVEL PROTOTYPE ALIGNMENT VIA CROSS-ATTENTION FOR FEW-SHOT REMOTE SENSING SEMANTIC SEGMENTATION

Mustafa Alawadi and Mansoor Fateh

(Received: 28-Jun.-2026, Revised: 8-Aug.-2026, Accepted: 10-Aug.-2026)

ABSTRACT

Analyzing remote sensing imagery relies heavily on accurate pixel-level classification to interpret complex geographical scenes. However, training robust deep-learning models demands densely annotated datasets, which are expensive and time-consuming to create. Few-shot semantic segmentation (FSS) provides a practical alternative by adapting to new categories from limited labeled examples. Despite success in natural images, direct application to remote sensing faces unique challenges, such as massive scale variations, strong visual similarities among densely packed objects and severe background interference. Existing FSS methods typically rely on masked average pooling for prototype extraction, which discards critical spatial geometries and fine-grained details. While some approaches attempt multi-scale or global-local modeling, they lack effective, decoupled alignment between support prototypes and query features, limiting their ability to simultaneously capture broad scene context and precise object boundaries in cluttered aerial views. To address these limitations, we propose DLPANet, a novel dual-level prototype alignment network centered on Prototype-Guided Spatial Attention (PGSA). Our framework constructs global and semantic prototypes from hierarchical backbone features, enabling simultaneous modeling of scene context and fine-grained details. Dual cross-attention modules then align these prototypes with decoupled global and semantic query feature maps, facilitating dynamic and precise feature enhancement that better distinguishes adjacent objects and suppresses background ambiguity. A learnable decoder integrates the enriched features, optimized via Dice loss to handle extreme class imbalance. Evaluated on the iSAID-5i benchmark, DLPANet achieves 36.12% and 41.05% mIoU in 1-shot and 5-shot settings, respectively, consistently outperforming state-of-the-art methods. These results demonstrate that our decoupled dual cross-attention mechanism provides superior prototype-query alignment compared to prior global-local strategies. https://github.com/mansoorfateh83-lab/dlpanet.

KEYWORDS

Few shot, Remote sensing, Segmentation, Prototype alignment, Cross-attention.

1. INTRODUCTION

Remote-sensing (RS) imagery analysis supports many essential applications, from environmental monitoring and urban planning to natural disaster management [1]-[2]. Semantic segmentation is a key part of this process, because it provides the pixel-level categorization needed to truly understand complex geographical scenes [3][4][5]. While deep-learning models perform exceptionally well on these tasks, their success depends heavily on massive, densely annotated datasets [6]-[7]. Creating these pixel-perfect annotations is an expensive and time-consuming bottleneck. Few-shot Semantic Segmentation (FSS) offers a practical solution to this problem, allowing models to identify and segment new object categories using only a few labeled support images [8]-[9].

Although FSS has proven highly effective for natural images, applying it directly to RS imagery is challenging [10]. Aerial images naturally contain extreme scale variations and densely packed objects that look very similar, such as confusing small cars with larger vehicles or ships. Furthermore, the appearance of objects can change drastically depending on the broader aerial context. These unique traits cause standard FSS models to struggle; they often fail to draw precise object boundaries and tend to mistakenly highlight irrelevant parts of the complex geographical background [11].

Researchers have developed specialized FSS frameworks to tackle these domain-specific hurdles. Early foundational models, like Scale-Aware Detailed Matching (SDM [12]), used multiple prototypes and a scale-aware focal loss to penalize large, well-segmented objects, forcing the network to focus on harder, smaller instances [12]. Other architectures, like R^2Net, introduced global rectification and decoupled registration to stop inaccurate prototype activation and separate foreground features from background

M. Alawadi and M. Fateh (Corresponding Author) are with Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran. Emails: mus.al.awadi@gmail.com and mansoor_fateh@shahroodut.ac.ir

clutter [13]. More recent work has pushed these ideas further. For example, some methods build stronger semantic connections by combining magnitude-based pruning with cross-matching and self-matching modules [6]. At the same time, newer frameworks, like SEMPNet, adapt vision-foundation models to turn segmentation into a precise mask classification task [14]. While these advancements certainly improve localization, a major flaw remains. Most current methods rely on masked average pooling to extract condensed object prototypes. This compression technique naturally throws away the spatial geometries and fine-grained semantic details needed for accurate segmentation. Additionally, simple cosine similarity matching between prototypes and query features is usually not strong enough to handle the complex backgrounds and extreme appearance changes found in aerial top-down views. Because of this, we need a better way to align semantic features and handle their interactions.

To move beyond the limitations of single-prototype matching and standard spatial-attention mechanisms in few-shot segmentation, we present a novel framework tailored for remote sensing imagery, built upon a dual-stream Prototype-Guided Spatial Attention (PGSA) strategy. A weight-sharing hierarchical backbone extracts multi-stage features from both support and query images. Rather than relying on a single compressed prototype, we construct two complementary support representations: a semantic prototype isolating target object features and a global prototype capturing broader spatial context. Concurrently, the query feature maps are decoupled across backbone stages into corresponding global and semantic feature maps. To establish dynamic spatial alignment, we apply a dual PGSA mechanism, where the semantic and global prototypes serve as compact semantic query anchors to compute spatially adaptive attention weights across the query key/value maps. The resulting contextually enriched vectors are merged back into the query feature space and processed by a learnable decoder to produce the final segmentation mask. By applying dual-stream PGSA to remote sensing FSS, our architecture effectively addresses severe background clutter and target scale variations.

The primary contributions of this research are established across several key aspects.

1) We propose a novel few-shot semantic-segmentation framework for remote-sensing imagery that leverages dual-level feature representations, capturing both global context and fine-grained semantic details.
2) We design a dual-stream Prototype-Guided Spatial Attention (PGSA) strategy tailored for remote sensing few-shot segmentation, which aligns support prototypes with query features at both global context and fine-grained semantic levels to effectively suppress background interference.
3) We design a learnable decoder that effectively integrates attention-enhanced features to produce accurate and dense segmentation masks.
4) We conduct extensive experiments on the iSAID-5i benchmark, demonstrating that the proposed method consistently outperforms existing state-of-the-art approaches.

The remainder of this paper is organized as follows. Section 2 reviews related work on semantic segmentation in remote sensing, general few-shot semantic segmentation and attention-based prototype learning. Section 3 presents the proposed DLPANet in detail, including the problem definition, overall framework, backbone network, dual-level prototype and query-feature extraction, dual cross-attention mechanism and the learnable decoder with Dice loss. Section 4 describes the experimental setup, reports quantitative and qualitative results on the benchmark and provides comprehensive ablation studies. Finally, Section 5 concludes the paper and discusses potential future directions.

2. RELATED WORK

This section reviews the literature relevant to few-shot semantic segmentation in remote-sensing imagery. We first discuss general semantic-segmentation techniques tailored for remote sensing data. We then review prominent advances in few-shot semantic segmentation for natural images. Subsequently, we focus on domain-specific methods developed for few-shot segmentation in aerial and remote-sensing scenarios. Finally, we highlight recent works on attention mechanisms and prototype-learning strategies that are most closely related to our proposed approach.

2.1 Semantic Segmentation in Remote Sensing

Semantic segmentation has become a fundamental task in remote-sensing image analysis, enabling pixel-level understanding of complex aerial scenes. Early deep-learning approaches were built upon

Fully Convolutional Networks (FCNs) [15], followed by encoder-decoder architectures, such as U-Net [16] and SegNet [17]. To better capture multi-scale context, DeepLab models [18]-[19] introduced atrous convolutions and spatial pyramid pooling.

Transformer-based architectures, such as SETR [20], have demonstrated strong capability in modeling global dependencies. In the remote-sensing domain, recent works increasingly integrate attention mechanisms and multi-scale feature fusion to address large spatial variations. For instance, attention-driven and sparse kernel-based segmentation architectures have been proposed to improve feature representation for high-resolution imagery [21]. Additionally, recent studies continue to explore multimodal and hybrid learning strategies to enhance segmentation robustness under complex sensing conditions [22].

2.2 Few-shot Semantic Segmentation

Prototype-based learning has become the dominant paradigm in FSS, where class-specific representations (prototypes) are extracted from support images and matched with query features.

2.2.1 General Few-shot Semantic Segmentation

Prototype-based learning is a dominant paradigm, where class representations are extracted from support images and matched with query features. PANet [23] introduced prototype-alignment regularization, while PFENet [24] incorporated prior masks for improved guidance.

Subsequent works improved representation learning and feature interaction. ASGNet [25] introduced adaptive prototypes and HSNet [26] modeled dense correspondences using hypercorrelation volumes. More recent developments have explored transformer-based designs and enhanced matching strategies.

In recent years, several studies further extended FSS by integrating additional modalities and learning paradigms. For example, text-guided few-shot segmentation frameworks incorporate textual prompts and graph reasoning to enhance semantic discrimination [27]. Similarly, diffusion-based data augmentation strategies have been proposed to synthesize novel samples and mitigate data scarcity in few-shot settings [28].

2.2.2 Few-shot Semantic Segmentation in Remote Sensing

Adapting FSS to remote-sensing imagery introduces challenges, such as extreme scale variation, dense object distribution and domain shifts. Early adaptations directly applied natural-image methods, but achieved limited success.

To address these issues, specialized approaches have been proposed. SDM [12] introduced scale-aware matching and multi-prototype learning, while R^2Net [13] leveraged global rectification and decoupled feature registration. SEMPNet [14] further reformulated the task as a mask-classification problem using foundation-model priors.

Recent works have significantly advanced this field. For instance, learnable-prototype frameworks introduce adaptive prototype representations using generative-modeling techniques to better capture intra-class variation [29]. Generalized few-shot segmentation has also been explored to handle both base and novel classes simultaneously, providing a more realistic evaluation setting [30]. Furthermore, cross-matching and self-matching mechanisms combined with network pruning have been proposed to improve feature correspondence and efficiency [6].

In addition, knowledge distillation and semi-supervised learning have been integrated in recent works to leverage unlabeled data and improve generalization under extremely limited annotations [31]. These approaches highlight a shift toward hybrid learning paradigms that combine supervision, self-supervision and data synthesis.

2.3 Attention Mechanisms and Prototype Learning

Attention mechanisms, particularly those introduced in the Transformer architecture [32], have become central to modern segmentation frameworks. Cross-attention enables dynamic interaction between support and query features, improving semantic alignment and robustness.

Despite their effectiveness, traditional prototype representations based on masked average pooling often

discard spatial information. Recent works attempt to address this limitation by introducing learnable prototypes, attention-guided refinement and multi-level feature aggregation. Notably, recent studies emphasize combining cross-attention with enhanced prototype learning to better capture both global context and fine-grained structures [6], [29].

Despite substantial progress, existing methods still struggle with complex spatial relationships, background clutter and severe domain-specific variations in remote-sensing imagery. Recent advances increasingly focus on integrating attention mechanisms, generative modeling and hybrid learning strategies. However, effective semantic alignment between support and query features remains an open challenge, motivating the proposed dual cross-attention framework.

Although existing methods have made notable progress, most still rely on simple masked average pooling for prototype extraction, which discards important spatial and fine-grained semantic details. Moreover, their feature-alignment strategies are often insufficient to handle the complex backgrounds and scale variations inherent in aerial imagery. To overcome these limitations, we propose DLPANet, which builds upon the prototype-based paradigm by introducing dual-level (global and semantic) prototypes and a powerful dual cross-attention mechanism for more effective and decoupled feature alignment between support and query images.

3. PROPOSED METHOD

3.1 Problem Definition

In the context of few-shot semantic segmentation for remote-sensing imagery, the primary objective is to accurately predict a dense, pixel-wise classification mask for a given query image utilizing a highly restricted set of annotated support images. Formally, the task is established over a training dataset $\mathcal{D}_{\text{train}}$ and a disjoint testing dataset $\mathcal{D}_{\text{test}}$, ensuring that the target semantic categories evaluated during the testing phase remain entirely unobserved during network optimization. Under a standard 1-way K-shot configuration, a learning episode is constructed by sampling a support set comprising K image-mask pairs, denoted as $\mathcal{S} = \left\{\left(x_s^{(i)}, y_s^{(i)}\right)\right\}_{i=1}^{K}$, alongside a single query image x_q. The support masks $y_s^{(i)} \in \{0,1\}^{H\times W}$ provide the spatial ground-truth for the novel target category. The underlying algorithmic goal is to extract and align optimal semantic concepts from the support set $\mathcal{S}$ to accurately generate a predicted segmentation mask $\hat{y}_q \in \mathbb{R}^{H\times W}$ for the query image x_q.

3.2 Overview

To address the severe scale variations and dense background interference inherent in top-down aerial imagery, a dual-pathway architecture driven by targeted Prototype-Guided Spatial Attention (PGSA) is established. As shown in Figure 1, the framework symmetrically processes the support and query images through a weight-sharing convolutional backbone to extract multi-level hierarchical representations.

Rather than relying on a singular pooled vector, the extracted support features are bifurcated to formulate both a global prototype and a semantic prototype, thereby capturing distinct spatial and contextual abstractions. Correspondingly, the query features are processed into decoupled global and semantic spatial-feature maps. A specialized cross-attention module is subsequently applied independently to both the global and semantic streams. Within this module, the condensed prototypes are utilized to query the flattened spatial-query feature maps, generating highly refined, attention-driven vectors. These enhanced vectors are spatially expanded, concatenated with the original deep query-feature maps and processed by a learnable decoder to yield the final segmentation mask.

3.3 Backbone Network

A pre-trained ResNet-50 architecture is employed as the foundational backbone network to provide robust, multi-scale hierarchical feature extraction. Both the support image and the query image are passed through this feature extractor, with the network weights strictly shared between the two branches to guarantee feature alignment in the embedding space. To comprehensively capture low-level spatial contours and high-level abstract semantics, intermediate feature maps are extracted from multiple stages of the ResNet-50 architecture. Specifically, outputs originating from Block 1 through Block 4 are extracted and subsequently concatenated along the channel dimension. A 1×1 convolutional layer is

then applied to the concatenated tensor to reduce the channel dimensionality, ensuring computational efficiency while preserving the critical spatial and semantic information required for precise dense prediction tasks.

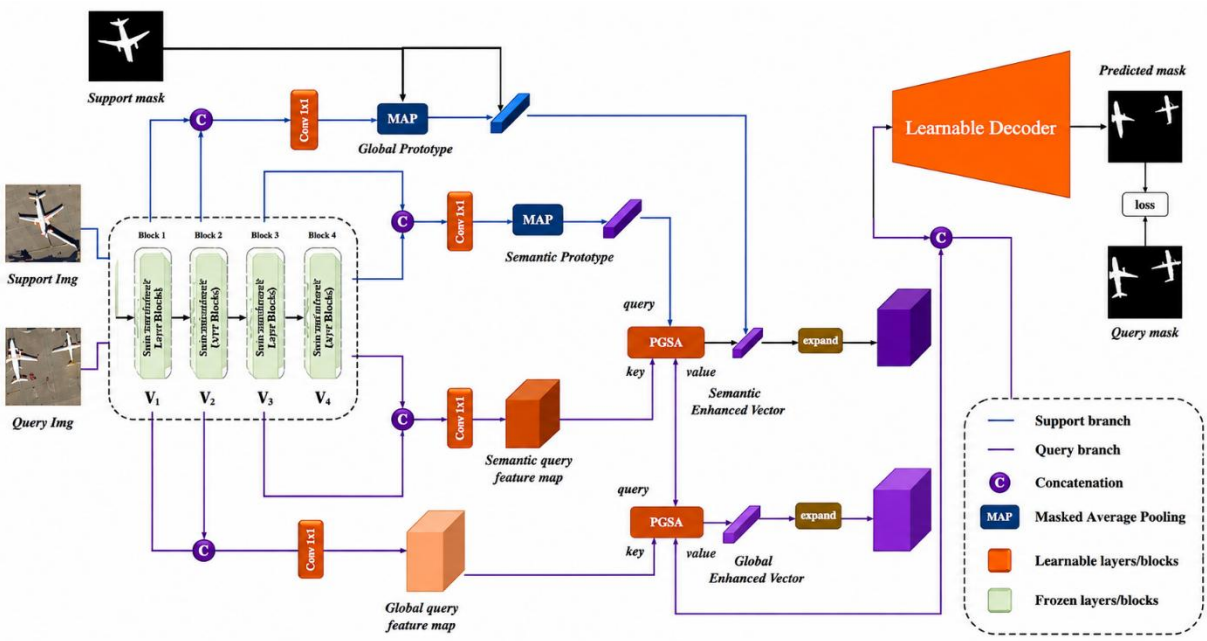


Figure 1. The overview of the proposed method.

3.4 Dual-level Prototype and Query Feature Extraction

To mitigate the degradation of fine-grained geometric details typical of single-prototype compression algorithms, a dual-representation strategy is implemented. The dimensionality-reduced support feature maps are explicitly processed to isolate targeted cues. To prevent background interference, the spatially down-sampled ground-truth support mask is applied across both support-extraction pathways, suppressing extraneous background clutter prior to prototype generation.

Specifically, Masked Average Pooling (MAP) is applied to the foreground-isolated deep support-feature maps (Blocks 3-4) to generate a focused, compact semantic prototype with dimensions of 1×64. Concurrently, a broader global prototype of identical dimensions (1×64) is extracted by applying masked average pooling to the projected early-stage support features (Blocks 1-2), capturing wider geometric and structural arrangements of the target object while excluding irrelevant background regions.

In symmetry, the query-feature maps are decoupled across fundamentally different backbone stages to capture complementary levels of contextual abstraction. Deep backbone stages (Blocks 3-4), which naturally possess a large receptive field and abstract categorical information, are projected *via* a 1×1 convolution $\phi_s(\cdot)$ to generate the semantic query-feature map. Concurrently, early backbone stages (Blocks 1-2), which preserve fine spatial geometries and low-level structural boundaries, are projected *via* an independent 1×1 convolution $\phi_c(\cdot)$ to formulate the global query-feature map.

Here, both $\phi_s(\cdot)$ and $\phi_c(\cdot)$ are parameterized as standard 1×1 2D convolutions (kernel size $= 1\times1$, stride $= 1$, padding $= 0$), each followed by a Batch Normalization layer and a ReLU non-linear activation function. These 1×1 convolutions perform channel-wise alignment from their respective backbone feature dimensions into a unified latent dimension ($d = 64$) without altering the distinct spatial resolutions or receptive fields established by the hierarchical backbone.

3.5 Prototype-guided Spatial Attention (PGSA)

To enable fine-grained and dynamic alignment between support prototypes and query features, DLPANet incorporates a dual Prototype-Guided Spatial Attention (PGSA) mechanism. Rather than relying on static cosine similarity or unweighted feature matching, PGSA computes an adaptive, content-dependent spatial weight map to aggregate query features relative to the support anchor.

As illustrated in Figure 2, we describe the semantic stream for clarity (the global stream follows an identical procedure). Given the semantic prototype $\mathbf{p}_s \in \mathbb{R}^{1\times 64}$ extracted from the support set, we treat it as a single prototype query anchor. Meanwhile, the corresponding semantic query feature map $\mathbf{F}_s^q \in \mathbb{R}^{64\times H'\times W'}$ is flattened into a sequence of spatial tokens $\mathbf{X}_s \in \mathbb{R}^{4096\times 64}$ (where $H' \times W' = 64 \times 64$).

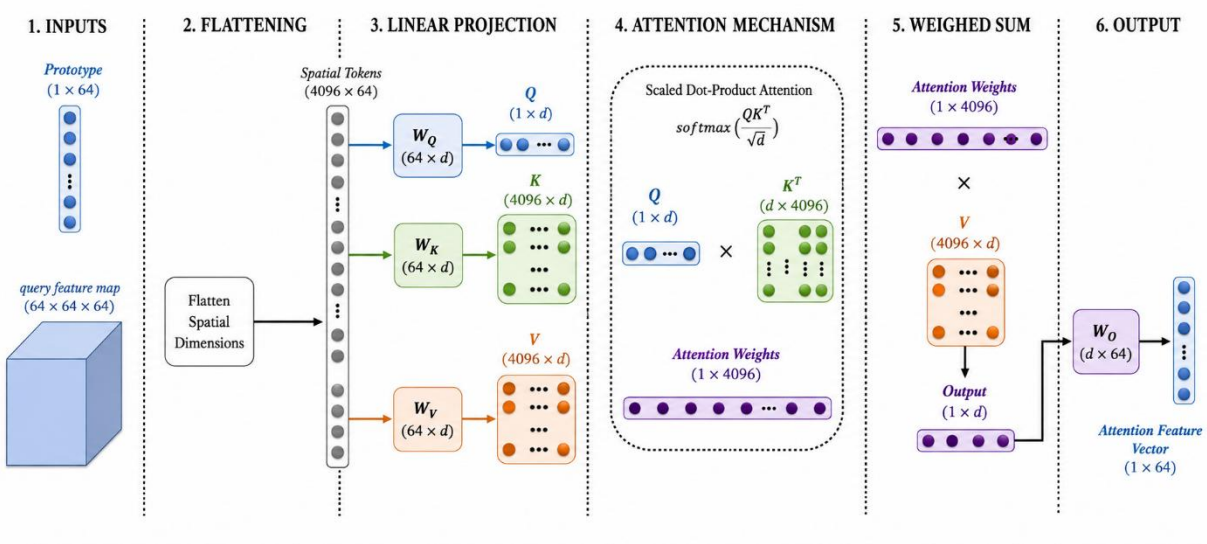


Figure 2. Overview of the proposed Prototype-Guided Spatial Attention (PGSA) module.

Learnable linear projection matrices $W_Q, W_K, W_V \in \mathbb{R}^{64\times d}$ (with $d = 64$ in our implementation) transform the inputs into query, key and value matrices:

$$\mathbf{Q} = \mathbf{p}_s W_Q, \mathbf{K} = \mathbf{X}_s W_K, \mathbf{V} = \mathbf{X}_s W_V \tag{1}$$

where $\mathbf{Q} \in \mathbb{R}^{1\times d}$ and $\mathbf{K}, \mathbf{V} \in \mathbb{R}^{4096\times d}$.

The spatial attention weights are computed *via* scaled dot-product attention:

$$A = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right) \in \mathbb{R}^{1\times 4096} \tag{2}$$

These weights represent the relevance of each spatial location in the query image relative to the support prototype anchor. The attended feature vector is subsequently obtained by:

$$\mathbf{O} = A\mathbf{V} \in \mathbb{R}^{1\times d} \tag{3}$$

Finally, a linear projection $W_O \in \mathbb{R}^{d\times 64}$ produces the enhanced semantic feature vector $\mathbf{z}_s \in \mathbb{R}^{1\times 64}$.

Mathematically, operating with a single query token $\mathbf{Q} \in \mathbb{R}^{1\times d}$ formulates a content-based spatial-attention pooling operation. While single-anchor attention pooling and cross-attention mechanisms have been explored in general few-shot segmentation architectures (e.g., BAM [33] and CyCTR [34]), the key distinction in DLPANet lies in how this operation is structured. Rather than applying attention over a single combined feature stream or performing dense token-to-token matching, our design operates in a decoupled dual-pathway fashion.

Specifically, the PGSA process is performed independently across both the semantic and global streams, yielding distinct enhanced feature vectors $\mathbf{z}_s$ and $\mathbf{z}_g$. By decoupling these streams, the model simultaneously attends to fine-grained object-specific details (semantic stream) and broader spatial geometry (global stream) at their native abstraction levels. This dual-stream alignment effectively mitigates background interference and resolves scale ambiguities that standard single-stream attention matching fails to address.

3.6 Learnable Decoder and Loss Function

Following the PGSA phase, the semantic and global attention feature vectors are spatially expanded to match the exact spatial resolution of the intermediate query-feature maps. The expanded global vector,

the expanded semantic vector and the rich query feature maps are concatenated along the channel dimension to form a comprehensive, contextually enriched feature tensor $\mathbf{Z}_{\text{cat}} \in \mathbb{R}^{C_{\text{in}} \times H' \times W'}$.

To reconstruct fine morphological details from this aggregated tensor, our learnable decoder employs a 3-stage progressive up-sampling block. Each stage consists of:

1) A bilinear up-sampling layer with a scale factor of 2 to enlarge feature map dimensions.
2) A standard 3×32D convolutional layer (padding = 1, stride = 1) to refine spatial cues and eliminate upsampling artifacts.
3) A Batch Normalization (BN) layer followed by a ReLU non-linear activation function.

Specifically, the channel dimension is progressively reduced across the three stages ($C_{\text{in}} \rightarrow 128 \rightarrow 64 \rightarrow 32$). Finally, a 1×1 convolution layer followed by a Sigmoid activation produces the final single-channel binary prediction map $\hat{\mathbf{Y}} \in \mathbb{R}^{1 \times H \times W}$ matching the original input-image resolution.

To optimize the network parameters and accurately handle extreme class imbalances frequently encountered in high-resolution remote sensing imagery, the Dice loss function is utilized. By calculating the spatial overlap between the predicted probability mask p and the binary ground-truth mask g, the Dice loss directly maximizes the intersection over union. The optimization objective is formulated as:

$$L_{\text{Dice}} = 1 - \frac{2\sum_{i=1}^{N}(p_i \cdot g_i) + \epsilon}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} g_i + \epsilon} \tag{4}$$

where N denotes the total number of pixels, p_i represents the predicted probability at pixel i, g_i indicates the corresponding binary ground-truth label and ϵ is a small smoothing constant added to prevent division by zero. To ensure numerical stability and exact experimental reproducibility, the smoothing constant is set to $\epsilon = 1 \times 10^{-5}$ in all experiments. This formulation ensures that the network optimization remains highly sensitive to the morphological boundaries of complex geographical targets, strictly penalizing spatial misalignments during the episodic meta-training phase.

4. EXPERIMENTAL RESULTS

4.1 Dataset

To thoroughly evaluate the effectiveness and generalization capability of the proposed DLPANet, we conduct experiments on two widely used few-shot semantic segmentation benchmarks: iSAID- 5^i for remote-sensing imagery and PASCAL- 5^i for natural images.

The iSAID-5^i benchmark is derived from the original iSAID dataset [35], which consists of 2,806 high-resolution aerial images with 655,451 annotated object instances across 15 geospatial categories. Following the standard cross-validation protocol [12], the 15 categories are divided into three folds (Fold 0, 1 and 2). In each evaluation, the model is meta-trained on the base classes from two folds and tested on the five novel classes of the remaining fold. Detailed class partitions are provided in Table 1.

For natural-image evaluation, we employ the PASCAL- 5^i benchmark [36]-[37]. This dataset is constructed from the PASCAL VOC 2012 dataset with additional dense annotations from the SDS dataset. It contains approximately 10,000 images spanning 20 object categories. Following the standard protocol, these categories are split into four folds ($i \in \{0,1,2,3\}$). In each fold, the model is meta-trained on 15 base classes and evaluated on the remaining 5 novel classes.

Experiments on both remote sensing (iSAID- 5^i) and natural image (PASCAL-5^i) domains allow us to validate the robustness and versatility of our dual cross-attention framework across different data distributions and challenges.

Table 1. Class partitions for the iSAID-5^i dataset across three folds.

# Fold	Novel Classes				
0	Ship (C_1)	Storage tank (C_2)	Baseball diamond (C_3)	Tennis court (C_4)	Basketball court (C_5)
1	Ground track field (C_6)	Bridge (C_7)	Large vehicle (C_8)	Small vehicle (C_9)	Helicopter (C_{10})
2	Swimming pool (C_{11})	Roundabout (C_{12})	Soccer ball field (C_{13})	Plane (C_{14})	Harbor (C_{15})

4.2 Implementation Details

The feature-extraction process is anchored by ResNet-50 backbone architectures, which is initialized with weights pre-trained on the ImageNet dataset to leverage established feature representations [13]. The meta-training phase is executed over 20 epochs utilizing a mini-batch size of 8. The base learning rate is established at 0.05. To enhance the generalization capabilities of the model and simulate diverse aerial perspectives, comprehensive data-augmentation techniques are applied to the training set. These techniques include random horizontal flipping, Gaussian blurring, random rotation spanning from -10° to 10° and random scaling operations ranging from a factor of 0.5 to 2.0. Furthermore, all raw input images are uniformly cropped to a spatial resolution of 512 × 512 pixels prior to being processed by the network. To facilitate exact experimental reproducibility, our complete PyTorch implementation and pre-trained weights are publicly available at https://github.com/mansoorfateh83-lab/dlpanet.

To verify the statistical significance and stability of the reported performance gains, all experiments for the proposed model are independently trained and evaluated across three distinct random seeds (100, 202 and 42). The final performance metrics are reported as Mean ± Standard Deviation across all runs.

4.3 Evaluation Metrics

To quantitatively evaluate the segmentation performance of the proposed DLPANet, we adopt two widely used metrics: the mean Intersection over Union (mIoU) and the Foreground-Background Intersection over Union (FB-IoU).

Mean Intersection over Union (mIoU): This serves as our primary evaluation metric. It computes the Intersection over Union (IoU) for each novel class independently and then takes the average across all target classes in the evaluation fold. The mIoU is defined as:

$$\mathrm{mIoU} = \frac{1}{C}\sum_{i=1}^{C}\frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i + \mathrm{FN}_i} \tag{5}$$

where C is the number of novel classes in the current fold and TP_i, FP_i and FN_i denote the number of true positive, false positive and false negative pixels for the i-th novel class, respectively. This metric provides a balanced and strict assessment by heavily penalizing errors on challenging categories.

Foreground-Background Intersection over Union (FB-IoU): As a complementary class-agnostic metric, FB-IoU evaluates the model's ability to distinguish objects of interest from the background without regard to specific semantic categories. All novel classes are merged into a single "foreground" category. The FB-IoU is computed as:

$$\mathrm{FB-IoU} = \frac{\mathrm{TP}_{\mathrm{fg}}}{\mathrm{TP}_{\mathrm{fg}} + \mathrm{FP}_{\mathrm{fg}} + \mathrm{FN}_{\mathrm{fg}}} \tag{6}$$

where $\mathrm{TP}_{\mathrm{fg}}$, $\mathrm{FP}_{\mathrm{fg}}$ and $\mathrm{FN}_{\mathrm{fg}}$ represent the aggregated true positives, false positives and false negatives for the combined foreground regions across all novel classes.

These two metrics together offer a comprehensive view of both per-class segmentation accuracy and overall foreground localization capability, which is particularly important in complex remote-sensing scenes with severe background interference.

4.4 Comparison with State-of-the-Art

The segmentation efficacy of the proposed dual cross-attention architecture is benchmarked against several prominent few-shot semantic segmentation (FSS) frameworks on the iSAID-5^{i} dataset. This dataset, characterized by its high-resolution aerial imagery and significant scale variations, serves as a rigorous testbed for evaluating the robustness of our dual-prototype generation and cross-attention alignment mechanism.

As detailed in Table 2, our proposed model achieves state-of-the-art results across both backbones. Specifically, with the ResNet-50 backbone, our architecture attains a mean mIoU of $36.12 \pm 0.28\%$ in the 1-shot setting and $41.05 \pm 0.32\%$ in the 5-shot setting, representing a substantial improvement over recent strong baselines, such as SFAPNet [43] and GlocalDualNet [44]. The small standard deviations

($\leq 0.35\%$) across three independent random seeds confirm that the performance gains are statistically significant and represent a reliable signal rather than stochastic noise. The consistent performance gains across all folds underscore the ability of the dual-prototype approach to capture more comprehensive semantic information than traditional single-prototype methods.

The superior localization capability of our method is further validated by the FB-IoU metrics reported in Table 3. FB-IoU provides a category-agnostic evaluation of the model's ability to distinguish foreground objects from complex backgrounds. The proposed approach reaches an FB-IoU of 66.30% and 69.25% for 1-shot and 5-shot scenarios, respectively. These results surpass existing methods, like GlocalDualNet [44], indicating that the dual cross-attention mechanism effectively mitigates foreground-background ambiguity-a common challenge in remote-sensing imagery due to diverse object orientations and cluttered environments.

Table 2. Performance comparison on iSAID -5^i in mIoU under 1-shot and 5-shot settings.

Backbone	Method	1-shot				5-shot			
		Fold-0	Fold-1	Fold-2	Mean	Fold-0	Fold-1	Fold-2	Mean
VGG-16	SDM [12]	24.52	21.01	16.31	20.61	26.73	19.97	26.10	24.27
	NERTNet [38]	25.78	20.01	19.88	21.89	38.43	24.21	28.99	30.54
	DML [39]	30.99	14.60	19.05	21.55	34.03	16.38	26.32	25.58
	TBPN [40]	27.86	12.32	18.16	19.45	32.79	16.28	24.27	24.45
	R ^{2} Net [13]	35.27	19.93	24.63	26.61	42.06	23.52	30.06	31.88
	CSCANet [41]	33.26	20.44	25.98	26.56	40.08	24.15	38.00	34.08
	DSMFNet [42]	30.23	23.35	28.87	27.48	40.73	25.92	37.05	34.56
	SFAPNet [43]	37.12	22.82	26.02	28.65	44.31	26.71	37.06	36.03
	GlocalDualNet [44]	36.40	20.51	27.59	28.17	42.64	23.55	34.74	33.64
ResNet-50	SDM [12]	27.96	21.99	27.82	25.92	28.50	25.23	31.07	28.27
	NERTNet [38]	34.93	23.95	28.56	29.15	44.83	26.73	37.19	36.25
	DML [39]	32.96	18.98	26.27	26.07	33.58	22.05	29.77	28.47
	TBPN [40]	29.33	16.84	25.47	23.88	30.98	20.42	28.07	26.49
	R ^{2} Net [13]	41.22	21.64	35.28	32.71	46.45	25.80	39.84	37.36
	CSCANet [41]	42.30	24.17	36.50	34.32	47.85	30.04	40.32	39.40
	DSMFNet [42]	39.03	24.82	39.57	34.47	43.69	30.81	42.32	38.94
	PGSL [45]	40.23	24.92	37.34	34.17	45.32	29.71	46.80	40.61
	SFAPNet [43]	43.48	25.03	37.18	35.23	48.35	29.63	48.41	42.13
	FSS-TIs [46]	-	-	-	24.51	-	-	-	31.21
	MSDNet [47]	44.35	28.71	35.16	36.07	48.59	29.01	40.64	39.41
	GlocalDualNet [44]	42.54	24.04	38.15	34.91	47.70	26.51	41.62	38.11
	Ours	**45.10±0.25**	**24.95±0.31**	**38.30±0.28**	**36.12±0.28**	**50.15±0.35**	**29.40±0.29**	**43.60±0.32**	**41.05±0.32**

Table 3. Performance comparison on iSAID-5 ^{i} in FB-IoU utilizing the ResNet-50 backbone.

Method	1-shot	5-shot
SDM [12]	59.58	59.90
NERTNet [38]	59.97	64.45
TBPN [40]	57.34	58.63
R^2 Net [13]	63.86	66.18
CSCANet [41]	63.56	66.32
DSMFNet [42]	63.52	66.31
SFAPNet [43]	64.47	66.75
GlocalDualNet [44]	65.05	66.97
Ours	**66.30**	**69.25**

A detailed class-wise breakdown of the mIoU performance under the 1-shot setting is provided in Table 4. Our model demonstrates leading performance in a majority of the 15 categories (C1-C15). Notably, the model shows significant improvements in challenging classes, such as C1, C2 and C13, which often involve high intra-class variance or small spatial extents. While established methods, like CSCANet [41], exhibit strengths in specific categories, like C12, our dual-prototype approach maintains a more balanced and higher overall mean performance. This suggests that the integration of hierarchical features

and cross-attention alignment provides a more generalized representation, ultimately leading to more reliable few-shot segmentation in diverse aerial scenes.

Table 4. Class-wise performance comparison in mIoU under the 1-shot setting with the ResNet50 backbone.

Methods	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	Mean
VGG-16 Backbone																
TBPN [40]	22.03	39.75	20.80	42.80	13.94	10.41	6.87	16.54	4.38	23.41	5.68	23.66	22.13	24.63	14.72	19.45
R^2 Net [13]	37.82	45.16	26.17	45.30	21.81	24.11	14.38	30.92	12.21	18.03	18.66	25.02	29.64	31.95	17.87	26.61
CSCANet [41]	36.21	43.88	26.01	43.39	16.81	21.80	15.84	26.65	10.58	27.33	9.05	41.67	32.19	31.01	15.97	26.56
SFAPNet [43]	41.48	49.92	31.52	39.95	22.93	26.23	24.23	33.37	13.06	17.21	10.77	27.93	38.58	32.68	19.92	28.65
ResNet-50 Backbone																
TBPN [40]	25.36	41.28	30.67	32.88	16.48	13.48	9.74	27.88	12.52	20.56	11.12	34.31	23.57	40.36	17.98	23.88
R^2 Net [13]	46.87	49.06	30.70	52.86	26.62	24.31	17.25	31.25	13.67	21.73	24.88	46.07	42.29	42.07	21.08	32.71
CSCANet [41]	45.96	47.83	36.62	57.99	23.10	21.27	23.45	29.47	11.98	34.28	18.69	59.39	37.45	46.80	20.17	34.32
Ours	**51.45**	**52.80**	**35.25**	**56.10**	**29.90**	**28.10**	**20.55**	**34.90**	**16.85**	**24.35**	**27.80**	**49.50**	**45.60**	**45.40**	**23.20**	**36.12**

Table 5. Performance on PASCAL-5^i in terms of mIoU and FB-IoU. Numbers in bold represent the best performance.

Backbone	Methods	1-shot						5-shot					
		fold0	fold1	fold2	fold3	mean	FB-IoU	fold0	fold1	fold2	fold3	mean	FB-IoU
ResNet50	PANet [23]	44.0	57.5	50.8	44.0	49.1	-	55.3	67.2	61.3	53.2	59.3	-
	PGNet [48]	56.0	66.9	50.6	50.4	56.0	69.9	57.7	68.7	52.9	54.6	58.5	70.5
	PFENet [24]	61.7	69.5	55.4	56.3	60.8	73.3	63.1	70.7	55.8	57.9	61.9	73.9
	HSNet [26]	64.3	70.7	60.3	60.5	64	76.7	70.3	73.2	67.4	67.1	69.5	80.6
	CWT [49]	56.3	62.0	59.9	47.2	56.4	-	61.3	68.5	68.5	56.6	63.7	-
	CyCTR [34]	65.7	71.0	59.5	59.7	64.0	-	69.3	73.5	63.8	63.5	67.5	-
	NTRENet [38]	65.4	72.3	59.4	59.8	64.2	77.0	66.2	72.8	61.7	62.2	65.7	78.4
	ABCNet [50]	62.5	70.8	57.2	58.1	62.2	74.1	64.7	73.0	57.1	59.5	63.6	74.2
	SRPNet [51]	62.8	69.3	55.8	58.1	61.5	-	64.3	70.3	55.1	60.5	62.6	-
	QGPLNet [52]	56.95	68.99	60.1	54.98	60.25	-	61.78	70.96	69.56	58.26	65.14	-
	DRNet [53]	66.1	68.8	61.3	58.2	63.6	76.9	69.2	73.9	65.4	65.3	68.5	81.6
	QGNet [54]	63.4	69.4	55.1	58.4	61.6	-	64.7	71.0	53.6	61.6	62.8	-
	Ours	**64.9**	**69.1**	**61.5**	**61.7**	**64.3**	**77.1**	**70.4**	**73.1**	**68.1**	**67.2**	**69.6**	**81.8**

To further validate the generalization capability of DLPANet beyond aerial imagery, we evaluate its performance on the widely used PASCAL-5^i benchmark for natural images. As shown in Table 5, our method achieves state-of-the-art performance under the ResNet-50 backbone in both 1-shot and 5-shot settings.

Specifically, DLPANet attains a mean mIoU of 64.3% in the 1-shot setting, outperforming previous best methods, such as DRNet and NTRENet. More importantly, it achieves the highest FB-IoU of 77.1%, indicating superior foreground-background separation. In the 5-shot setting, our approach further improves to 69.6% mIoU and 81.8% FB-IoU, surpassing all competing methods.

These strong results on the PASCAL-5^i dataset demonstrate that the proposed dual-level prototype alignment via cross-attention is not limited to the unique challenges of remote-sensing imagery. The consistent superiority across both domains highlights the effectiveness and versatility of our dual cross-attention mechanism in establishing robust support-query feature correspondences.

4.5 Computational Efficiency Analysis

To demonstrate the practical suitability of DLPANet for deployment in high-resolution remote-sensing

applications, we evaluate its computational complexity against leading baseline models. Table 6 summarizes the Number of Parameters (#Params in Millions), Computational Complexity (FLOPs in Giga-FLOPs), Inference Speed (Frames Per Second, FPS) and Mean mIoU (%).

As presented in Table 6, DLPANet achieves superior segmentation accuracy (36.12% mIoU) while maintaining remarkable computational efficiency. Specifically, by utilizing single-token prototype query anchors and lightweight 1 × 1 linear projections rather than dense sequence-to-sequence cross-attention layers, DL-PANet operates with only 24.72M parameters and 20.5 GFLOPs. Furthermore, DLPANet achieves an inference rate of 34.2 FPS, outperforming all baseline models in speed and proving that its performance gains are achieved without imposing prohibitive computational burdens.

Table 6. Computational efficiency comparison under the 1-way 1-shot setting with ResNet-50 backbone.

Method	Params (M) (↓)	FPS (↑)	FLOPs (G) (↓)	Mean mIoU (%) (↑)
PANet	64.31	27.6	24.8	23.13
CANet	69.01	25.3	26.5	21.15
SCL	58.61	31.2	22.3	29.58
PFENet	57.51	33.8	21.7	28.80
DMML	70.31	29.7	24.6	24.31
SDM	76.01	27.3	22.9	25.92
R^2 Net	51.76	32.5	21.3	32.71
GlocalDualNet	60.15	30.8	25.7	34.91
Ours	**24.72**	**34.2**	**20.5**	**36.12**

4.6 Ablation Study

To rigorously validate the individual contributions of the architectural components within the proposed framework, an extensive ablation study is conducted. All ablation experiments are evaluated utilizing the ResNet-50 backbone under the 1-way 1-shot setting on the iSAID-5^{i} benchmark, unless otherwise specified. The mean Intersection over Union (mIoU) across all three folds is reported to ensure a comprehensive evaluation.

4.6.1 Effectiveness of the Dual PGSA

To demonstrate the necessity of our dual-pathway feature alignment, the proposed DLPANet is systematically deconstructed and compared against a defined baseline model.

To ensure a fair and controlled comparison, the baseline architecture utilizes the exact same dual-branch ResNet-50 feature extractor and 1×1 linear projection heads (Q_{sem} from Blocks 3-4, Q_{glob} from Blocks 1-2) as DLPANet. However, it completely omits the proposed Prototype-Guided Spatial Attention (PGSA) modules. Instead, it extracts support prototypes *via* standard Masked Average Pooling (MAP) and computes spatial query alignment using conventional cosine similarity matching, followed by a standard 1×1 convolution to generate foreground predictions.

As detailed in Table 7, this baseline model achieves a mean mIoU of 28.51%. Integrating the semantic prototype pathway alongside its dedicated PGSA module (PGSA_{sem} significantly enhances the delineation of object boundaries, raising the performance to 32.40% mIoU (Variant A). Conversely, utilizing only the global prototype pathway with its dedicated attention module ($\text{PGSA}_{\text{glob}}$ yields 31.85% mIoU (Variant B), demonstrating its effectiveness in capturing broader scene context. The optimal segmentation accuracy of 36.12% mIoU is exclusively attained when both global and semantic PGSA pathways are fully integrated. This confirms that holistic scene understanding and targeted semantic alignment function as mutually reinforcing mechanisms.

4.6.2 Analysis of Feature Projection Strategy for Query Branch

To evaluate whether explicitly expanding the spatial receptive field during query projection improves feature decoupling, we compare our lightweight 1 × 1 convolutional projection against an Atrous Spatial Pyramid Pooling (ASPP) module with multi-scale dilation rates {1,6,12} applied to the global query stream.

Table 7. Ablation study on the core architectural components. $PGSA_{sem}$ and $PGSA_{glob}$ denote the semantic and global PGSA modules, respectively.

Model	$PGSA_{sem}$	$PGSA_{glob}$	Fold-0	Fold-1	Fold-2	Mean mIoU
Baseline	×	×	31.20	20.15	34.18	28.51
Variant A	✓	×	38.65	23.40	35.15	32.40
Variant B	×	✓	37.90	22.85	34.80	31.85
Ours	✓	✓	**45.10**	**24.95**	**38.30**	**36.12**

As shown in Table 8, replacing the 1×1 projection with an ASPP module leads to a performance drop across all folds, reducing the mean mIoU from 36.12% to 34.85%. In remote-sensing imagery, where objects such as small vehicles, ships and narrow bridges occupy small spatial extents, ASPP's aggressive multiscale dilation introduces feature over-smoothing and boundary distortion. In contrast, utilizing linear 1×1 channel projection directly on the multi-stage features extracted from the hierarchical backbone preserves crisp spatial geometries while allowing the cross-attention mechanism to dynamically align global and semantic contexts without introducing spatial artifacts.

Table 8. Comparison of different projection strategies for the global query branch.

Projection Strategy	Fold-0	Fold-1	Fold-2	Mean mIoU
ASPP Module	43.80	24.10	36.65	34.85
1×1 Conv Projection	45.10	24.95	38.30	36.12

4.6.3 Analysis of the Loss Function

The choice of loss function plays a crucial role in few-shot remote sensing semantic segmentation due to the severe class imbalance between small foreground objects (e.g., small vehicles, ships and storage tanks) and vast background regions. Standard losses often struggle with this imbalance, leading to biased predictions toward the dominant background class.

Table 9 compares three different optimization strategies. Cross-Entropy (CE) Loss, while simple, tends to produce overconfident predictions and is highly sensitive to class imbalance. Scale-Aware Focal Loss improves performance by focusing more on hard examples, yet it still falls short in effectively managing the extreme foreground-background imbalance common in aerial imagery. In contrast, the proposed Dice Loss directly optimizes the overlap between predicted and ground-truth masks. By emphasizing region-level similarity rather than pixel-wise classification, Dice Loss more effectively penalizes false positives in complex background areas while improving boundary precision. This leads to the best overall performance of 36.12% mean mIoU, demonstrating its superior suitability for high-resolution remote-sensing scenarios.

Table 9. Effect of different optimization loss functions on segmentation accuracy.

Loss Function	Fold-0	Fold-1	Fold-2	Mean mIoU
Cross-Entropy (CE) Loss	42.15	23.10	35.10	33.45
Focal Loss ($\gamma = 2$)	43.80	24.15	36.45	34.80
Dice Loss (Ours)	45.10	24.95	38.30	36.12

4.6.4 Impact of the Decoder Architecture

After enriching query features through dual cross-attention, effectively decoding and upsampling these features into a high-resolution segmentation mask is essential. Table 10 evaluates different decoding strategies.

A simple 1×1 convolution decoder, which directly projects the concatenated features, yields a relatively low performance of 34.10% mIoU due to its limited capacity to recover spatial details lost during feature extraction. Replacing it with an Atrous Spatial Pyramid Pooling (ASPP) module improves multi-scale context modeling, boosting the result to 35.30% mIoU. However, our learnable decoder, composed of

progressive convolutional upsampling layers with skip connections, achieves the highest accuracy of 36.12% mIoU. This decoder better restores fine spatial geometries and sharp object boundaries by gradually refining the feature representations, which proves particularly beneficial for delineating small and densely packed objects in aerial scenes.

Table 10. Comparison of different decoding strategies for the final mask generation.

Decoder Strategy	Fold-0	Fold-1	Fold-2	Mean mIoU
Simple Conv (1 × 1)	43.05	23.50	35.75	34.10
ASPP Module	44.25	24.10	37.55	35.30
Learnable Decoder (Ours)	45.10	24.95	38.30	36.12

4.6.5 Multi-shot Fusion Strategy

In the more practical K-shot setting ($K > 1$), effectively aggregating information from multiple support images is critical for robust prototype construction. To justify our architectural design choices, we internally implemented and evaluated two distinct fusion strategies under the 5-shot setting, as presented in Table 11.

The naive Mean Pooling approach simply averages the prototypes extracted from the five support samples. While straightforward, this method treats all support images equally and fails to exploit complementary information across samples, resulting in a lower performance of 38.50% mean mIoU (a baseline ablation result not included in the main comparison table). In contrast, our proposed Attention-based Fusion strategy routes all support samples through the dual cross-attention modules. This allows the network to dynamically weigh and integrate the most relevant features from different support examples according to their spatial and semantic correspondence with the query image. As a result, the full DLPANet model achieves a substantial improvement of 2.55% in mean mIoU (reaching 41.05%, as reported in Table 2), confirming that adaptive, attention-guided fusion is significantly more effective than simple averaging for multi-shot scenarios.

Table 11. Ablation study on the impact of multi-shot feature fusion strategies under the 5-shot setting.

Fusion Strategy (5-Shot)	Fold-0	Fold-1	Fold-2	Mean mIoU (%)
Mean Pooling (Averaging)	47.90	26.40	41.20	38.50
Attention-based Fusion	50.15	29.40	43.60	41.05

4.6.6 Latent Channel Dimension

To justify our choice of the latent channel dimension d (obtained *via* the initial 1×1 projection convolutions and maintained across the PGSA modules), we evaluated performance and parameter efficiency across $d \in \{32,64,128\}$. As reported in Table 12, setting $d = 32$ leads to a notable drop in segmentation performance (32.40% mean mIoU), as a lower-dimensional embedding lacks the capacity to retain fine-grained spatial and semantic details.

Conversely, increasing d to 128 yields marginal performance gains (36.25% *vs.* 36.12% mIoU) while substantially increasing the trainable parameter count and memory footprint. Setting $d = 64$ achieves an optimal trade-off between representational accuracy and computational efficiency, securing superior performance with minimal parameter overhead.

Table 12. Ablation study on the latent channel dimension d under the 5-shot setting.

Dimension (d)	Fold-0	Fold-1	Fold-2	Mean mIoU (%)	Decoder Parameters (M)
$d = 32$	41.20	21.10	34.90	32.40	1.85M
$d = 64$	45.10	24.95	38.30	36.12	2.42M
$d = 128$	45.25	25.10	38.40	36.25	4.18M

4.7 Visualization

Figure 3 presents qualitative comparisons that highlight the incremental contributions of the proposed components on the iSAID-5 ^{i} benchmark. The baseline model, which relies on standard masked average pooling and simple feature matching, produces fragmented predictions with severe background interference and poor boundary delineation - common failure cases in existing few-shot remote-sensing methods.

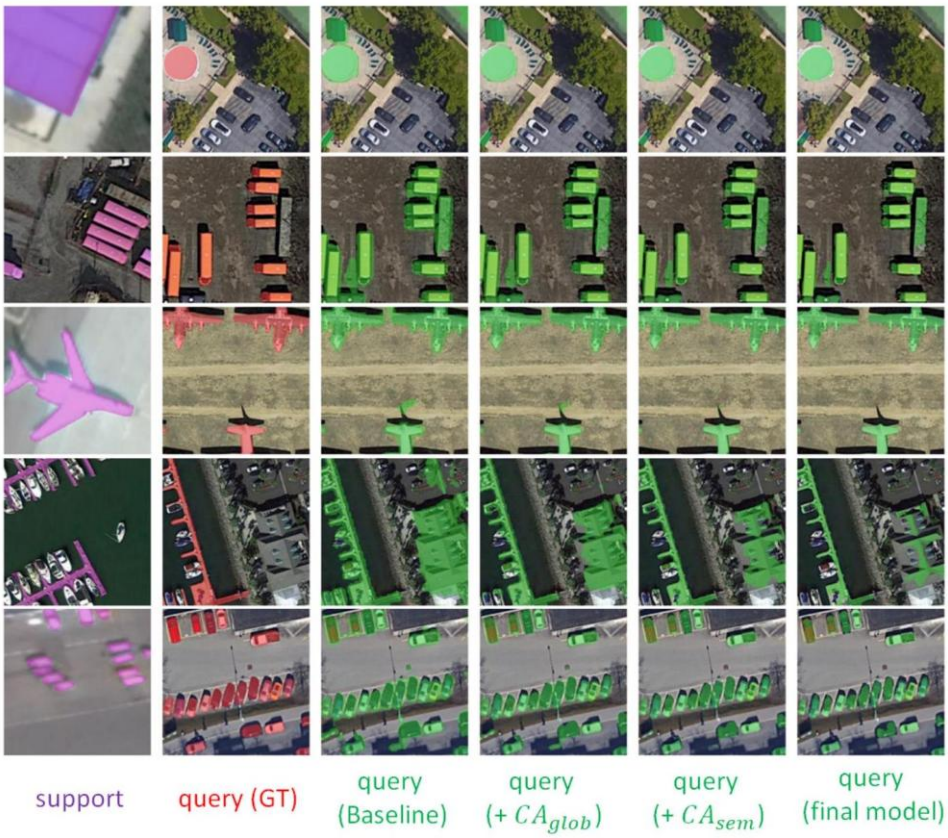


Figure 3. One-shot segmentation predictions on the iSAID-5 ^{i} benchmark. From left to right: support image with mask, query image with ground truth, baseline prediction, baseline with CA_{glob} only, baseline with CA_{sem} only and the full model combining both modules.

When only the global cross-attention module (CA_{glob}) is added, the model demonstrates a strong capability in suppressing large-scale background clutter and better capturing the overall scene context, resulting in cleaner object localization. Conversely, incorporating only the semantic cross-attention module (CA_{sem}) significantly improves fine-grained detail preservation and sharpens object boundaries, effectively reducing boundary overshoot and intra-class confusion between visually similar adjacent objects.

The full DLPANet model, which integrates both global and semantic streams through dual cross-attention, achieves the most accurate segmentation. It successfully handles complex challenges that previous methods struggle with: distinguishing densely packed objects with similar appearances, maintaining precise boundaries across varying scales and robustly suppressing intricate geographical background interference. The final predictions closely align with the ground truth, preserving fine spatial geometries while maintaining contextual coherence. These visual results clearly demonstrate that the decoupled dual cross-attention mechanism provides superior prototype-query alignment capabilities unavailable in conventional single-prototype approaches, enabling more reliable few-shot semantic segmentation in challenging aerial environments.

5. CONCLUSION

In this study, a novel architecture driven by dual PGSA was proposed to address the persistent challenges of extreme scale variations and dense background interference in few-shot remote-sensing semantic

segmentation. To overcome the inherent limitations of conventional single-prototype representations, dual-level representations-specifically, a global prototype and a semantic prototype-were constructed to simultaneously capture broad scene-level context and preserve fine-grained spatial details. These prototypes were actively aligned with decoupled query features through specialized cross-attention modules, which facilitated dynamic feature interaction and ensured refined spatial correspondences.

Extensive evaluations were conducted on the iSAID-5i benchmark, demonstrating that the proposed framework consistently established new state-of-the-art performance standards across existing methodologies. Specifically, exceptional mean Intersection over Union scores alongside superior localization capabilities, as evidenced by the category-agnostic Foreground-Background Intersection over Union metrics, were achieved under both 1-shot and 5-shot evaluation configurations. The critical contributions of the core architectural components, including the dual cross-attention pathways, the integration of the Dice loss function to manage severe class imbalances and the deployment of a learnable decoder for precise dense mask generation, were strictly validated through rigorous ablation studies. Ultimately, it is demonstrated by these quantitative findings that comprehensive hierarchical feature alignment combined with specialized interaction mechanisms significantly enhances the overall robustness and boundary accuracy of semantic segmentation frameworks operating within complex aerial domains.

REFERENCES

[1] B.-Y. Li et al., "Exploring Efficient Open-Vocabulary Segmentation in the Remote Sensing," Proc. of the AAAI Conf. on Artificial Intelligence, volume 40, pp. 5982-5991, 2026.

[2] K. Cheikh, R. Aitahcene, A. Toumi and Z. Hammoudi, "Fusion of Deep Learning Architectures for Enhanced Target Recognition on Sar Images," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 9, no. 4, pp. 347-359, 2023.

[3] H. Bi et al., "ViRefSAM: Visual Reference-guided Segment Anything Model for Remote Sensing Segmentation," IEEE Trans. on Geoscience and Remote Sensing, vol. 64, pp. 5606720- 5606720, 2026.

[4] M. Gholami, M. Fateh and A. Fateh, "Text-enhanced Semantic Segmentation *via* Contrastive Language-image Pre-training Guided Multi-modal Feature Fusion with Feature Refinement Approach," Int. Journal of Engineering, vol. 39, no. 6, pp.1422-1437, 2026.

[5] A. Saber et al., "A Lightweight Multiscale Refinement Network for Gastrointestinal Disease Classification," Expert Systems with Applications, vol. 308, Article no. 131029, 2026.

[6] K. Amoafo et al., "Enhancing Few-shot Semantic Segmentation in Remote Sensing through Magnitude-based Pruning," Signal Processing: Image Communication, vol. 143, Article no. 117492, 2026.

[7] G. Hcini and I. Jdey, "Privacy Aware Malaria Detection: U-net Model with K-anonymity for Confidential Image Analysis," Jordan. J. of Computers and Information Tech. (JJCIT), vol. 11, no. 1, pp.85-99, 2025.

[8] A. Fateh, M. R. Mohammadi and M. R. Jahed-Motlagh, "Adapting SAM with a triple-prompt strategy for one-shot Semantic Segmentation," Neurocomputing, vol. 681, p. 133301, 2026.

[9] N. Singhal et al., "Study of Recent Image Restoration Techniques: A Comprehensive Survey," Jordanian J. of Computers and Information Technology (JJCIT), vol. 11, no. 2, pp. 211-237, 2025.

[10] Y. Shi et al., "Twostage Fine-tuning of Large Vision-language Models with Hierarchical Prompting for Few-shot Object Detection in Remote Sensing Images," Remote Sensing, vol. 18, no. 2, p. 266, 2026.

[11] S. Zhang, X. Zhang, Y. Xu and K. Wang, "Few-shot Object Detection on Remote Sensing Images Based on Decoupled Training, Contrastive Learning and Self-training," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 19, pp. 3983-3997, 2026.

[12] X. Yao et al., "Scale-aware Detailed Matching for Few-shot Aerial Image Semantic Segmentation," IEEE Trans. on Geoscience and Remote Sensing, vol. 60, Article no. 5611711, pp. 1-11, 2021.

[13] C. Lang et al., "Global Rectification and Decoupled Registration for Few-shot Segmentation in Remote Sensing Imagery," IEEE Trans. on Geosci. and Remote Sens., vol. 61, Art. no. 5617211, pp. 1-11, 2023.

[14] W. Ao, S. Zheng and Y. Meng, "SEMPNet: Enhancing Few-shot Remote Sensing Image Semantic Segmentation through the Integration of the Segment Anything Model," GIScience & Remote Sensing, vol. 61, no. 1, Article no. 2426589, 2024.

[15] J. Long, E. Shelhamer and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," Proc. of the IEEE Conf. on Computer Vision and Patt. Recog. (CVPR), pp. 3431-3440, Boston, USA, 2015.

[16] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," Proc. of the 18th Int. Conf. on Medical Image Computing and Computer-assisted Intervention (MICCAI 2015), pp. 234-241, Munich, Germany, October 5-9, 2015.

[17] V. Badrinarayanan et al., "SegNet: A Deep Convolutional Encoder-decoder Architecture for Image Segmentation," IEEE Trans. on Pattern Analy. and Machine Intelli., vol. 39, no. 12, pp. 2481-2495, 2017.

[18] L.-C. Chen, G. Papandreou, F. Schroff and H. Adam, "Rethinking Atrous Convolution for Semantic

Image Segmentation," arXiv preprint, arXiv: 1706.05587, 2017.
[19] L.-C. Chen et al., "Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation," Proc. of the Europ. Conf. on Comp. Vision (ECCV2018), vol. 11211, pp. 801-818, 2018.
[20] S. Zheng et al., "Rethinking Semantic Segmentation from a Sequence-to-sequence Perspective with Transformers," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 6881-6890, Nashville, USA, 2021.
[21] M. Fu et al., "LSKSANet: A Novel Architecture for Remote Sensing Image Semantic Segmentation Leveraging Large Selective Kernel and Sparse Attention Mechanism," Proc. of 2024 IEEE Int. Geoscience and Remote Sensing Symposium (the IGARSS 2024), pp. 8438-8441, Athens, Greece, 2024.
[22] H. Feng et al., "FTransDeepLab: Multimodal Fusion Transformer-based Deeplabv3+ for Remote Sensing Semantic Segmentation," IEEE Trans. on Geosci. and Remote Sens., vol. 63, Article no. 4406618, 2025.
[23] K. Wang et al., "PANet: Few-shot Image Semantic Segmentation with Prototype Alignment," Proc. of the 2019 IEEE/CVF Int. Conf. on Computer Vision, pp. 9197-9206, Seoul, S. Korea, 2019.
[24] Z. Tian et al., "Prior Guided Feature Enrichment Network for Few-shot Segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 2, pp. 1050-1065, 2020.
[25] G. Li et al., "Adaptive Prototype Learning and Allocation for Few-shot Segmentation," Proc. of the 2021 IEEE/CVF Conf. on Comp. Vision and Pattern Recog. (CVPR), pp. 8334-8343, Nashville, USA, 2021.
[26] J. Min, D. Kang and M. Cho, "Hyper-correlation Squeeze for Few-shot Segmentation," Proc. of the 2021 IEEE/CVF Int. Conf. on Computer Vision, pp. 6941-6952, Montreal, Canada, 2021.
[27] F. Yang, J. Chen, D. Luo and H. Lv, "Few-shot Segmentation of Remote Sensing Imagery with Textual Prompts," Int. Journal of Remote Sensing, vol. 47, no. 3, pp. 1040-1060, 2026.
[28] S. A. Immanuel, W. Cho, J. Heo and D. Kwon, "Tackling Few-shot Segmentation in Remote Sensing *via* Inpainting Diffusion Model," arXiv preprint, arXiv: 2503.03785, 2025.
[29] J. Wang et al., "Few-shot Semantic Segmentation on Remote Sensing Images with Learnable Prototype," IEEE Trans. on Geoscience and Remote Sensing, vol. 63, Article no. 5104411, 2025.
[30] C. Broni-Bediako et al., "Generalized Few-shot Semantic Segmentation in Remote Sensing: Challenge and Benchmark," IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1-5, 2024.
[31] H.-L. Jiang et al., "A Few-shot High-resolution Remote Sensing Image Semantic Segmentation Method," Scientific Reports, vol. 16, Article no. 15262, 2026.
[32] A. Vaswani et al., "Attention Is All You Need," Advances in Neural Inf. Process. Syst., vol. 30, 2017.
[33] C. Lang, G. Cheng, B. Tu and J. Han, "Learning What Not to Segment: A New Perspective on Few-shot Segmentation," Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 8057-8067, New Orleans, USA, 2022.
[34] G. Zhang, G. Kang, Y. Yang and Y. Wei, "Few-shot Segmentation *via* Cycle-consistent Transformer," Advances in Neural Information Processing Systems, vol. 34, pp. 21984-21996, 2021.
[35] S. W. Zamir et al., "iSAID: A Large-scale Dataset for Instance Segmentation in Aerial Images," Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition Workshops, pp. 28-37, 2019.
[36] M. Everingham, L. V. Gool, C. K. Williams, J. Winn and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," Int. J. of Computer Vision, vol. 88, pp. 303-338, 2010.
[37] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji and J. Malik, "Semantic Contours from Inverse Detectors," Proc. of the IEEE 2011 Int. Conf. on Computer Vision, pp. 991-998, Barcelona, Spain, 2011.
[38] Y. Liu et al., "Learning Non-target Knowledge for Few-shot Semantic Segmentation," Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 11573-11582, New Orleans, 2022.
[39] X. Jiang, N. Zhou and X. Li, "Few-shot Segmentation of Remote Sensing Images Using Deep Metric Learning," IEEE Geoscience and Remote Sensing Letters, vol. 19, Article no. 6507405, pp. 1-5, 2022.
[40] G. Puthumanaillam and U. Verma, "Texture Based Prototypical Network for Few-shot Semantic Segmentation of Forest Cover: Generalizing for Different Geographical Regions," Neurocomputing, vol. 538, Article no. 126201, 2023.
[41] G. Liang, F. Xie and Y.-R. Chien, "Class-aware Self-and Cross-attention Network for Few-shot Semantic Segmentation of Remote Sensing Images," Mathematics, vol. 12, no. 17, pp. 2761, 2024.
[42] X. Qi et al., "DSMF-Net: Dual Semantic Metric Learning Fusion Network for Few-shot Aerial Image Semantic Segmentation," IEEE J. of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 853-864, 2024.
[43] S. Wang, M. Meng, M. Chang and Z. Yang, "Learning Spatial-frequency Adaptive prototype for Remote Sensing Few-shot Segmentation," Knowledge-based Systems, vol. 349, Article no. 116464, 2025.
[44] H. Tang, Y. Jia, J. Cheng, Y. Mu, Y. Wu and X. Yao, "Glocaldualnet: Disentangling Scale and Representation for Few-shot Remote Sensing Segmentation," IEEE J. of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 19, pp. 8018-8030, 2026.
[45] Q. Wang, H. Jiang, J. Feng, G. Zhang and J. Yin, "Prototype-guided Structural Learning from Visual Foundation Model for Few-shot Aerial Image Semantic Segmentation," Proc. of the 2024 IEEE Int. Geoscience and Remote Sensing Symposium (IGARSS 2024), pp. 8433-8437, Athens, Greece, 2024.
[46] H. Ni, Q. Liu, X. Niu, D. Hong, L. Zhao and H. Guan, "Cross-domain Few-shot Segmentation *via*

Ordinary Differential Equations over Time Intervals," arXiv preprint, arXiv: 2509.01299, 2025.
[47] A. Fateh et al., "MSDNet: Multi-scale Decoder for Few-shot Semantic Segmentation *via* Transformer-guided Prototyping," Image and Vision Computing, vol. 162, p. 105672, 2025.
[48] C. Zhang, G. Lin, F. Liu, J. Guo, Q. Wu and R. Yao, "Pyramid Graph Networks with Connection Attentions for Region-based One-shot Semantic Segmentation," Proc. of the 2019 IEEE/CVF Int. Conf. on Computer Vision, pp. 9587-9595, Seoul, S. Korea, 2019.
[49] Z. Lu et al., "Simpler Is Better: Few-shot Semantic Segmentation with Classifier Weight Transformer," Proc. of the 2021 IEEE/CVF Int. Conf. on Computer Vision, pp. 8741-8750, 2021.
[50] Y. Wang, R. Sun and T. Zhang, "Rethinking the Correlation in Few-shot Segmentation: A Buoys View," Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 7183-7192, Vancouver, Canada, 2023.
[51] H. Ding, H. Zhang and X. Jiang, "Self-regularized Prototypical Network for Few-shot Semantic Segmentation," Pattern Recognition, vol. 133, Article no. 109018, 2023.
[52] Y. Tang and Y. Yu, "Query-guided Prototype Learning with Decoder Alignment and Dynamic Fusion in Few-shot Segmentation," ACM Trans. on Multimedia Computing, Communications and Applications, vol. 19, no. 2s, Article no. 84, pp. 1-20, 2023.
[53] Z. Chang et al., "DRNet: Disentanglement and Recombination Network for Few-shot Semantic Segmentation," IEEE Trans. on Circuits and Sys. for Video Tech., vol. 34, no. 7, pp. 5560-5574, 2024.
[54] W. Liu, Z. Wu, H. Ding, F. Liu, J. Lin, G. Lin and W. Zhou, "Exploiting Independent Query Information for Few-shot Image Segmentation," Displays, vol. 91, Article no. 103179, 2026.

ملخص البحث:

يعتمد تحليل صور الاستشعار عن بعد بشكلٍ كبيرٍ على التّصنيف الدّقيق لتفسير المشاهد الجغرافية المعقّدة. ومع ذلك، فإنّ تدريب نماذج تعلّم عميق قوية يتطلّب مجموعات بيانات ذات تعليقات توضيحية كثيفة (موسومة بدقّة)، وهي عملية مكلفة وتستغرق وقتاً طويلاً. وتوفّر تقنية التّجزئة الدّلالية باستخدام عينات قليلة بديلاً عملياً من خلال التّكيف مع فئات جديدة استناداً الى أمثلةٍ محدودةٍ وموسومة. ورغم النّجاح المحقّق في الصُّور الطّبيعية. فإنّ التّطبيق المباشر لهذه التّقنية في مجال الاستشعار عن بُعدٍ يواجه تحدّياتٍ فريدةً مثل التّباين الهائل في الأحجام، والتّشابه البصري القوي بين الكائنات المتراصّة بكثافة، والتّداخل الشّديد من الخلفية.

وتعتمد طرق التّجزئة الدّلالية باستخدام عينات قليلة عادةً على تقنية التّجميع المتوسّط المقنّع لاستخراج النّماذج الأولية، ممّا يؤدّي إلى فقدان معلوماتٍ هندسيةٍ مكانيةٍ حاسمةٍ وتفاصيلٍ دقيقةٍ. وفي حين تحاول بعض الأساليب استخدام النّمذجة متعدّدة المقاييس أو النّمذجة الشّاملة (العالمية والمحلية)، فإنّها تفتقر إلى آلية محاذاةٍ فعّالةٍ ومنفصلة بين نماذج الدّعم الأولية وميزات الاستعلام، مما يحدّ من قدرتها على إلتقاط سياق المشهد العامّ وحدود الكائنات الدّقيقة في آنٍ واحدٍ ضمن المشاهد الجوّية المزدحمة. ولمعالجة هذه القيود، نقترح شبكةً مبتكرة لمحاذاة النّماذج الأولية ثنائية المستوى ترتكز على آلية الانتباه المكاني الموجّه بالنّماذج الأولية. ويقوم إطار العمل المقترح ببناء نماذجٍ أوليةٍ عالميةٍ ودلاليةٍ انطلاقاً من ميزات هيكلية هرمية، مما يتيح نمذجة متزامنة لسياق المشهد والتّفاصيل الدّقيقة. بعد ذلك، تقوم وحدات الإنتباه المتقاطع المزدوج بمحاذاة هذه النّماذج الأولية مع خرائط ميزات الاستعلام العالمية والدّلالية المنفصلة، مما يسهّل تعزيز الميزات بشكل ديناميكي ودقيق، ويساعد بشكلٍ أفضل على التّمييز بين الكائنات المتجاورة والحدّ من غموض الخلفية. وأخيراً، يقوم مفكّك ترميز قابل للتّعلّم بدمج الميزات مع تحسين الأداء باستخدام دالة الخسارة للتّعامل مع مشكلة عدم التّوازن الشّديد بين الفئات. وعند تقييمها، تحقّق الشّبكة المقترحة تفوّقاً على أحدث الأساليب المتطوّرة في هذا المجال.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

FR-CNN: SOFTWARE FAULT PREDICTION USING A NOVEL HYBRID HIERARCHICAL MODEL

Mehrasa M. Jouybari, Alireza Tajary, Mansoor Fateh and Esmaeel Tahanian

(Received: 16-May-2026, Revised: 29-Jul.-2026, Accepted: 19-Aug.-2026)

ABSTRACT

Developing high-quality software products is directly related to the absence of unseen faults. The software fault prediction (SFP) methods have been considered to reduce the overall testing costs and discover probable faults. Various machine-learning and deep-learning algorithms have been used to predict faults. However, both deliver varying results regarding SFP. This research proposes a novel hybrid model (FR-CNN) based on a fine-tuned fully connected deep neural network, a random forest and a convolutional neural network. Harnessing the strengths of combining these techniques, the proposed model aims to enhance prediction performance while reducing overfitting, resulting in a robust framework. We conduct extensive empirical studies to illustrate the effectiveness of the proposed model in predicting faults on six projects of the BugHunter dataset. The empirical results demonstrate that FR-CNN achieves competitive fault-prediction performance while requiring lower computational complexity and training time than recent deep-learning approaches. In particular, it improves the best previously reported traditional machine-learning result on the BugHunter benchmark by up to 19.95% in F1-score.

KEYWORDS

Software-fault prediction, Class imbalance, Deep learning, Ensemble learning, Machine learning, Hyper-parameters.

1. INTRODUCTION

The size and complexity of software in various applications are increasing and software engineers are tasked with developing high-quality software products. The existence of faults is one of the main issues that threaten the software quality and reliability [1] [2]. A software fault indicates a defect or deficiency in the system that leads to the final product's failure [3, 4]. Hidden and unforeseen faults are unavoidable, as they grow within the proper platform facilitated by modern software [5]. Over 50% of the costs of software development can be attributed to the detection and elimination of faults, as well as the removal of adverse consequences resulting from system failures [6].

Software Fault Prediction (SFP) is an approach that estimates the probability of faults in software units without running and testing the code. This strategy can effectively contribute to the production of higher-quality software [7]. Anticipating and early elimination of possible faults in different phases of development provides higher levels of system quality and improves maintenance effort [8]. SFP process allocates more effort to the most faulty modules and pays attention to diminishing the overall cost of testing [4], [9].

Various machine learning (ML) algorithms including Logistic Regression (LR) [10], Support Vector Machine (SVM) [11], K-Nearest Neighbors (KNN) [12], Decision Tree (DT) [13], Genetic Algorithm (GA) [14], Random Forest (RF) [15], Naive Bayes (NB) [16] and neural network [17], have been studied and investigated for SFP [12]. Nevertheless, these methods produce varying results in SFP. Deep learning (DL), as a subfield of ML, in a supervised and unsupervised manner, has achieved commendable success in various domains, including speech recognition and image processing [18]. DL is a proper approach to dealing with complex problems [19], demonstrating exemplary performance across both large and small datasets [20]. The DL model automatically extracts essential information using the features in the known project and after learning knowledge from them, it is used to predict faults in the unknown project. This approach affects the fault-prediction process. Convolution Neural Network (CNN) is a popular architecture of neural networks in SFP, capable of acquiring semantic features from software metrics [18].

Recent studies have further demonstrated that hybrid deep-learning architectures have become one of the most promising directions for software-fault prediction. Abdu et al. [21] proposed a hybrid deep-

M. M. Jouybari (Corresponding Author), A. Tajary, M. Fateh and E. Tahanian are with Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran. Emails: {mehrasa.modanlou; tajary; mansoor_fateh; e.tahanian}@shahroodut.ac.ir

learning framework that fuses semantic and traditional software features to improve defect prediction performance. Kaliraj et al. [22] introduced a dynamic classification framework for software-fault prediction that improves prediction robustness through adaptive classification strategies. More recently, Fei et al. [23] presented a multivariate heterogeneous hybrid deep-learning architecture that combines multiple complementary learning mechanisms for software-defect prediction, while Malhotra and Singh [24] proposed a dual-attention hierarchical BiGRU architecture that enhances software-defect prediction through hierarchical feature learning and attention mechanisms. These recent developments indicate that integrating complementary machine-learning and deep-learning models is becoming an increasingly effective strategy for improving software-fault prediction performance, which strongly motivates the proposed FR-CNN architecture.

The efficiency of the SFP on ML and DL techniques is also influenced by different fault datasets. The majority of well-known datasets for SFP, such as PROMISE [25] and NASA MDP [26], have relatively small sizes. Neural networks, particularly deep neural networks, are data-hungry models that their ability to learn increases with a larger amount of training data [27]. In this research, the BugHunter dataset [28], a recently introduced fault dataset comprising a sufficient number of data instances, has been used to satisfy the data requirements of DL-based models. This dataset consists of 15 popular open-source Java projects widely used by software engineers.

Although BugHunter is a recently introduced large-scale software fault-prediction dataset that provides a valuable benchmark for evaluating software fault-prediction models, its adoption in the literature remains relatively limited compared with widely used benchmark datasets, such as PROMISE and NASA. Existing approaches primarily rely on individual machine-learning models, conventional ensemble techniques or standalone deep-learning architectures, each of which has inherent limitations in exploiting complementary feature representations. Consequently, comprehensive evaluations of advanced hierarchical hybrid machine-learning and deep-learning models for method-level software fault prediction on the BugHunter dataset remain scarce. Motivated by recent advances in machine learning and deep learning, together with the availability of the large-scale BugHunter dataset, this study investigates whether a hierarchical hybrid architecture can improve software-fault prediction performance. Accordingly, we propose the FR-CNN model, which integrates a fine-tuned FC-DNN, Random Forest and a CNN meta-learner to exploit the complementary strengths of machine learning and deep learning for method-level software-fault prediction on six selected BugHunter projects.

The objective of this study is to develop and evaluate the proposed FR-CNN model for method-level software-fault prediction on six selected BugHunter projects and to compare its performance with the developed traditional machine-learning baseline models and the implemented recent state-of-the-art approaches under identical experimental conditions. The main contributions of this paper are as follows:

- We apply the random over-sampling technique on the BugHunter dataset to address the class-imbalance issue and enhance the training accuracy on both fault classes.
- We propose a new Fully Connected Deep Neural Network (FC-DNN) structure for predicting faults in six projects of the BugHunter dataset. We construct and evaluate multiple architectures for each project, investigating the effect of diverse hyper-parameters on improving the model performance until we achieve the optimal one.
- We propose the FR-CNN, a novel hierarchical hybrid model that combines a fine-tuned FC-DNN and the best-performing ML algorithm (Random Forest) using a CNN-based meta-learner. We further investigate the influence of kernel size, batch size and number of filters on the model performance.
- The proposed FR-CNN overcomes the limitations of existing CNN-based, ensemble-learning and hybrid software fault-prediction approaches through two key architectural innovations: (a) a hierarchical stacking architecture in which a fine-tuned FC-DNN and Random Forest (RF) serve as complementary base learners—each model learns different patterns from the software metrics and our feature importance analysis (Sub-section 4.6.2) confirms that they focus on diverse aspects of the feature space; and (b) a one-dimensional CNN as the meta-learner, enabling the learning of a combination rule over the base-learner outputs. An ablation study (Sub-section 4.6.7) shows that this learned combination outperforms parameter-free averaging on average across the six projects and that the CNN performs comparably to alternative learnable meta-classifiers (DNN, Logistic Regression) while remaining substantially more lightweight than the base-learner networks that it

combines. To the best of our knowledge, this specific FR-CNN hierarchical combination has not previously been investigated for
method-level software-fault prediction on the BugHunter dataset.

- We conduct comprehensive empirical studies by comparing the proposed FR-CNN with seven traditional machine-learning models and five recent state-of-the-art approaches using prediction performance, training time, testing time and model complexity. The results demonstrate that FR-CNN achieves competitive prediction performance while providing a favorable balance between computational efficiency, model complexity and training stability.

The paper proceeds as follows: Section 2 reviews previous works and contains the activities carried out in the SFP. Section 3 presents the proposed method. Section 4 pertains to the results and addresses the research questions. Section 5 deals with threats to validity of the proposed approach. Section 6 provides conclusions and describes future work.

2. RELATED WORKS

Numerous studies are available in the literature to classify software components as faulty or non-faulty, leveraging different techniques. These techniques were applied to various datasets at different granularity levels. Kumar et al. [8] collected and categorized various ML techniques, including NB, Decision Tree Regression and LR, for predicting fault with several performance evaluation measures. French et al. [28] developed a novel dataset called "BugHunter" to predict software faults using 11 traditional ML techniques from the Weka library, including RandomForest, J48, RandomTree, DecisionTable, SimpleLogistic and Logistic algorithms. Among the results presented for this dataset, the best F-measure at the method level was 75.73%, obtained by the RF algorithm for the ANTLR v4 project.

Cynthia et al. [29] developed a Feature Transformation for Improved Software Bug Detection Models that utilize a genetic algorithm to find the weighted transformation of the extracted features and identify buggy and non-buggy classes using ML models. They've reported their findings at the Class granularity and applied the undersampling technique to balance the dataset. They concluded that their proposed method can effectively separate bug data from non-bug data in the low-dimensional space of the transformed features.

Reference [30] introduced the Static Execute After (SEA)-based metrics instead of traditional software metrics to predict software faults at the method level of the BugHunter dataset. We will compare our proposed method with this model to understand how well our approach predict faulty methods. This research utilized random under-sampling to address the imbalance of two bug classes. The experimental results on the most frequently ML algorithms illustrated that the proposed SEA metrics are even suitable to predict future bugs of the project independently, if sufficient learning data is available.

So far, several studies have been carried out for SFP using ensemble learning-based techniques [31] [32] [33] [34]. These techniques are utilized to surpass the accuracy threshold of individual learners and increase the model's global predictive performance [35] [36]. According to [15], EL models, which are created by combining multiple ML classifiers, perform significantly better than individual classifiers such as decision trees, naïve Bayes or a multilayer perceptron for the fault prediction. Laradji et al. [37] formed a software fault-prediction method using the Average Probability Ensemble learning (APE) module with seven basic classifiers to predict faulty components.

Wang et al. [38] presented a Multiple Kernel ensemble-learning (MKEL) approach using EL, considering a multiple kernel classifier to improve software module classification and prediction power with faults. The experimental results showed the effectiveness of the proposed model on 12 NASA datasets. Pandey et al. [39] proposed an effective BPDET model for SFP using EL and deep representation techniques. The proposed approach was carried out in two phases, one was DL and the other was two layers of EL (TEL). The results of experiments on 12 public NASA datasets showed that BPDET is stable and the combination of deep representation and EL is more efficient and effective than the base methods and other traditional methods.

Although conventional ensemble-learning methods have demonstrated significant improvements over individual classifiers, recent research has shifted toward hybrid deep-learning frameworks that combine complementary feature representations. Abdu et al. [21] integrated semantic and traditional software

metrics using a hybrid deep-learning architecture and demonstrated that feature fusion significantly improves software-defect prediction. Likewise, Kaliraj et al. [22] proposed a holistic software fault-prediction framework incorporating dynamic classification mechanisms to improve prediction robustness. These studies indicate that combining different learning paradigms provides more informative feature representations than relying on a single prediction model.

DL techniques using convolutional layers, multilayer perceptron and Long Short-Term Memory (LSTM) automatically extract semantic features of the dataset [40]. Thus, these techniques create potent models that perform significantly in diverse domains, such as natural-language processing, computer vision and speech recognition. Reference [41] used the DL technique for fault prediction and presented the Seml framework. In this technique, the long short-term memory model performed fault prediction by automatically learning the semantic information of programs. The evaluation results on 13 Java open-source projects showed that the proposed model performs more promisingly compared to the state-of-the-art fault-prediction approaches. Alenezi et al. [42] employed character n-gram embedding with deep learning to improve source-code representation for software-vulnerability detection, achieving better prediction performance than previous approaches.

The latest research continues this trend. Fei et al. [23] proposed a multivariate heterogeneous hybrid deep-learning model that combines multiple deep-learning components for software-defect prediction, demonstrating improved prediction accuracy on complex software projects. Similarly, Malhotra and Singh [24] developed a dual-attention hierarchical gated BiGRU architecture that improves software defect prediction through hierarchical recurrent feature learning and attention mechanisms. Samal [43] proposed a hybrid optimization-based fault-prediction approach integrating particle-swarm optimization with fuzzy time series modeling. Collectively, these studies demonstrate that hybrid architectures capable of learning complementary representations have become one of the dominant research directions in software fault prediction.

In our research, we aim to combine these techniques to harness the advantages of ML and DL approaches while leveraging the strength of the ensemble model.

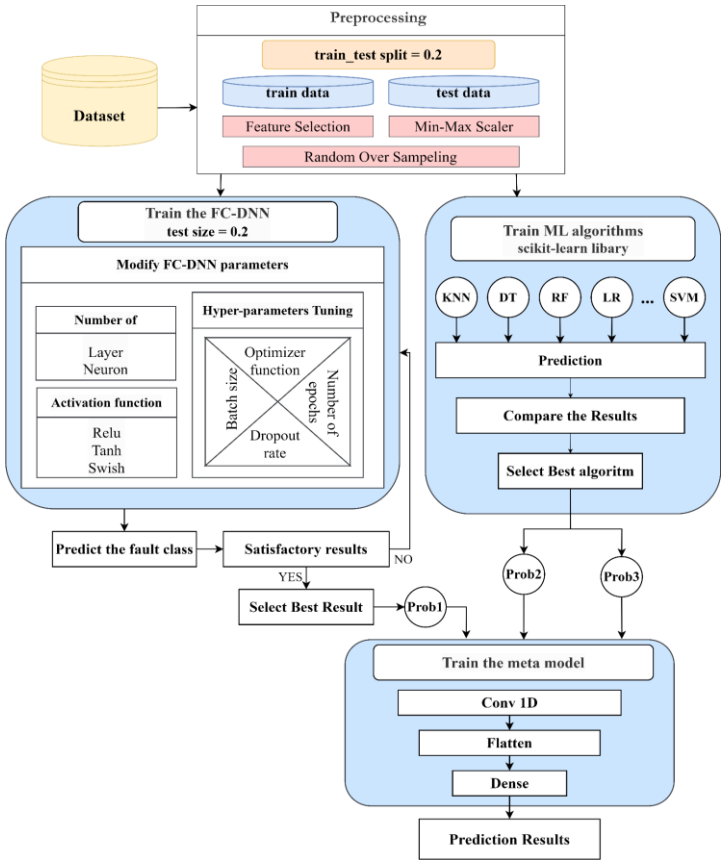


Figure 1. Overall workflow of the proposed hierarchical FR-CNN framework, illustrating the pre-processing stage, the FC-DNN and RF base learners and the CNN meta-learner for software-fault prediction.

3. PROPOSED APPROACH

This section introduces the proposed FC-DNN and FR-CNN models. The framework's overview is shown in Figure 1. To predict the number of faults at the method level of six Java projects from the BugHunter dataset, we employ the following steps. First, we conduct the pre-processing by choosing six datasets with various fault percentages, selecting and normalizing software metrics and addressing the class-imbalance problem. Second, we design multiple architectures for fault-prediction models using a DL structure and fine-tune them.

We also train various ML classifiers to be utilized in our hybrid model. Then, we develop an FR-CNN hybrid model by combining the optimized deep neural network and the best-performing ML model and feed them into a CNN structure. Our proposed approach includes two key phases: Data Pre-processing and Model Building.

3.1 Data Pre-processing

Preprocessing the dataset using different techniques improved the effectiveness of the model for learning [44]. To accomplish this task while strictly preventing data leakage, we applied a carefully ordered sequence of preprocessing steps. All data transformations and oversampling were performed using parameters learned exclusively from the training data after the train-test split, with the test data being transformed using the same parameters (e.g., min-max scaling), but never used to fit any transformations or oversampling, ensuring unbiased performance evaluation. The complete pre-processing workflow is as follows:

3.1.1 Feature Selection

The selected dataset contains a set of software metrics that describe information about each instance to predict the fault class at the method level. However, some of the included software metrics within the dataset, such as the 'Hash', 'LongName' and 'Migration14 Rules' metrics, do not possess sufficient value as reliable indicators for predicting faults [45]. Therefore, these metrics should be excluded, so that the selected metrics represent the original ones [46]. Additionally, we removed constant features (those with identical minimum and maximum values), as they provide no discriminative information. If the corresponding project at the method level is indicated by M and the fault class is denoted by Y, then the fault-prediction issue can be modeled according to Equation (1):

$$Y = f(M) \tag{1}$$

Some software metrics may have data with different scales, redundant features or useless information for prediction, which require pre-processing. Equation (2) shows the chosen features, where the M project being represented as a set of s_m, with each s_m denoting a software metric.

$$M = < sm_1, sm_2, \ldots, sm_n > \tag{2}$$

3.1.2 Null-value Verification

We verified that the dataset contains no null values to ensure data integrity before proceeding with further pre-processing steps.

3.1.3 Train-Test Split

To prevent data leakage, we first split the dataset into training and test sets using train_test_split from sklearn with a test size of 0.2. This ensures that the test data remains completely isolated and is never used for learning any transformations or model parameters throughout all subsequent pre-processing and model-building stages.

3.1.4 Binary Classification of the Output Class

In the selected dataset, the "Number of Bugs" software metric is considered as the label and exhibits a maximum of seven classes, which determine the count of faults within the output class. During the phase of prediction, the classification of the output class is considered as the presence or absence of faults for each individual sample. Hence, it is necessary to convert the label from the multi-class classification mode to the binary class classification. Therefore, instances characterized by the class label of 0 reflect the absence of faults, while instances associated with class labels ranging from 1 to 6 indicate the

presence of faults. Based on this, all labels have been modified to either 0 or 1. This conversion was applied to both training and test sets separately using the same mapping rule.

3.1.5 Data Normalization

Some ML techniques perform better on numerical datasets that have been scaled in a specific range. We performed the metric normalization on the dataset to ensure a fair balance in the range of data and equalize the effect of all dimensions. To achieve this, we applied the min-max scaler exclusively to all metrics of the training data, excluding the label. The same scaler fitted on the training data was then used to transform the test data. This was done using Equation (3), which rescaled the values of each software metric to a new range of 0 to 1.

$$\text{Normalization } (sm_i) = (sm_i - \min(sm)) / (\max(sm) - \min(sm)) \quad (3)$$

where sm_i indicates the corresponding software metric before applying normalization, max (sm) and min (sm) indicate the maximum and minimum values of sm_i and Normalization (sm_i) indicates the normalized values of sm_i. After normalization, the method M is defined according to Equation (4), which includes the set of sm_i' software metrics.

$$M = < sm_1', sm_2', \dots, sm_n' > \quad (4)$$

3.1.6 Addressing the Imbalance Issue

Unbalanced classes of the dataset reduce the classifier performance for the minority class. To overcome this, oversampling and undersampling techniques are applied to change the data space [47]. These techniques use strategies that either increase the instances of the minority class or reduce the instances of the majority class [48] to advance the final efficiency of the model [49]. In this article, the Random Over Sampling (ROS) technique was applied exclusively on the training instances to handle the imbalance distribution of the class labels, prevent the model from bias and enhance the performance of the learning process. The test data was not oversampled to maintain its original distribution and provide a realistic evaluation of model performance. Figure 2 and Figure 3 illustrate the count of two class instances in the BroadleafCommerce project, depicting the dataset's state prior to and following the usage of ROS on the training data. First, the number of instances was 826 for class 1 and 2941 for class 0. After applying ROS, the count of two class instances became equal.

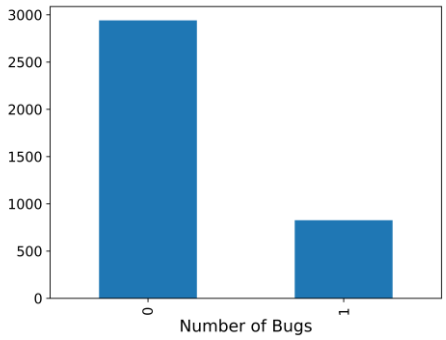


Figure 2. Class distribution of the BroadleafCommerce training data before applying Random oversampling (ROS), illustrating the original class imbalance.

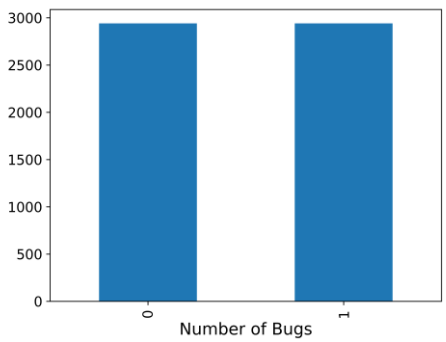


Figure 3. Class distribution of the BroadleafCommerce training data after applying Random oversampling (ROS), showing the balanced class distribution used for training.

It is important to emphasize that this sequential pre-processing workflow strictly prevents data leakage: the train-test split is performed before any transformations, the min-max scaler is fitted only on the training data and then applied to transform the train and test data and oversampling is applied exclusively to the training data. Consequently, no information from the test data is used during pre-processing or model training, ensuring an unbiased evaluation of model performance.

3.2 Model Building

3.2.1 Building the FC-DNN

The proposed FC-DNN, depicted in Figure 4, is built by a structure of dense and dropout layers. This fully connected feedforward deep neural network has an input layer, an output layer and between 1 and 4 hidden layers. There is a dropout layer after each layer except the last layer that prevents the model from overfitting by randomly dropping out the incoming and outgoing connections of units [50]. This regularization technique, combined with early stopping during training of the FR-CNN model, ensures robust generalization performance, as evidenced by the training and validation loss curves presented in Sub-section 4.6.4.

The model input is fixed according to the number of software metrics for each project which falls within the range of 56 to 59.

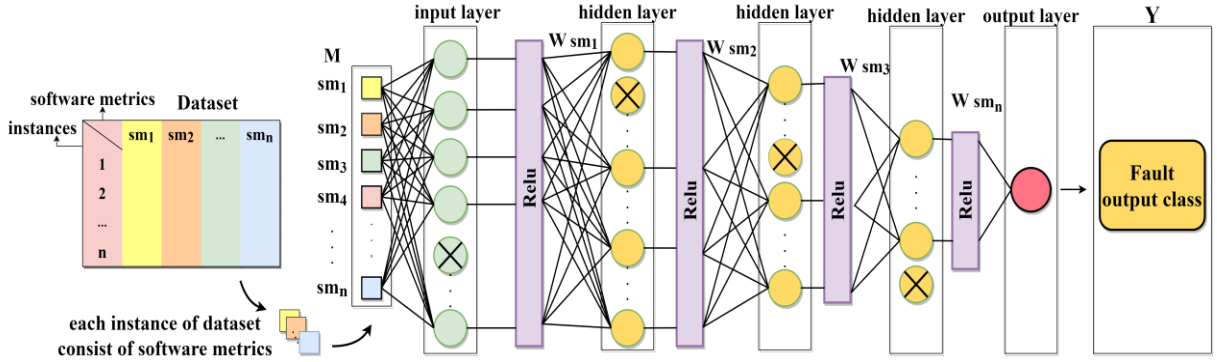


Figure 4. Architecture of the proposed FC-DNN base learner used for software-fault prediction.

The network output is considered in binary mode, with one neuron in the last layer. The model adjusts its weights using the optimizer function to reduce the error in the backpropagation stage [19]. The weights of layers are represented using $W_{sm_1}, W_{sm_2}, W_{sm_3}, \dots, W_{sm_n}$ in Figure 4. In the Result Section, we will conduct many experiments to design prediction models on each project, modify their parameters and record the best results. The impact of modifying parameters has an essential role in prediction performance [50]. The number of layers is a critical criterion in the architecture of the artificial neural networks and can influence the performance of the models [51]. We will modify and evaluate various hyper-parameters, different numbers of layers and activation functions to obtain the optimal design of the network and more accurate prediction models. Tuning the model hyper-parameters has been carried out by manipulating the number of epochs, batch sizes, dropout rates and optimizer functions until we reach satisfactory results. We will use three optimizer functions, Adam (Adaptive Moment Estimation), AdaGrad (Adaptive Gradients) and SGD (Stochastic Gradient Descent). We will utilize the Sigmoid activation function for the last layer of the network and test the ReLU, Tanh and Swish activation functions for other layers to choose the most suitable ones.

3.2.2 Building the FR-CNN Hybrid Model

In this Sub-section, the FR-CNN hybrid model is presented to improve the fault-prediction model's performance. The structure of the FR-CNN is depicted in Figure 1. This model is constructed by combining the FC-DNN, which is optimized with hyper-parameters and the RF, which is identified as the best-performing ML algorithm. Subsequently, a single meta-learner based on CNN is employed, comprising a convolutional layer, a flatten layer and a dense layer, serving as the final model. This model extracts semantic features from the combined features of both base models and determines the fault class for the entire hybrid model. To build the effective FR-CNN, we analyzed various

combinations of models. For instance, we investigated the potential of combining one or more ML classifiers with an FC-DNN serving as the base models, along with an ML, an FC-DNN or an CNN as a meta-learner. Through our evaluations, we found that the most effective approach for detecting software faults was achieved by combining an FC-DNN with an RF and putting a CNN at the end of it. Among various ML models that we established, the standalone RF exhibited superior performance in the hybrid model. To leverage its strengths, we combined it with a FC-DNN model as the base models. By combining these two models of different natures and distinct characteristics, we aimed to cover some of their individual weaknesses and enhance overall performance. In the proposed FR-CNN, the RF is constructed by aggregating the result of 50 decision trees. Given the existence of numerous outliers in the dataset under investigation, it is advantageous to adopt a clustering-oriented perspective during data analysis. Through the process of partitioning data among the decision trees and enabling independent decision-making by each tree, the RF algorithm exhibits a clustering-like behavior. Moreover, RF stands out as a robust EL-based algorithm that has an outlier-detection mechanism [52]. Additionally, we incorporated a fully connected DL-based algorithm into our methodology, leveraging its capability to automatically learn and capture distinguishing features and discriminative characteristics in the dataset. The utilization of this type of model has the potential for contributing towards the advancement of more accurate fault-prediction models [53]. In the Result Section, we will demonstrate the effectiveness of the presented FR-CNN on the BugHunter dataset. We also investigate the effect of different kernel sizes, batch sizes and numbers of filters on the performance of the FR-CNN. The workflow illustrated in Figure 1 and summarized in Algorithm 1 is mathematically formulated in Equations (5)–(8). Specifically, the formulation describes how the FC-DNN and Random Forest (RF) base learners generate prediction probabilities that are concatenated into meta-features and subsequently processed by the CNN meta-learner to produce the final software-fault prediction.
Let $x_i \in \mathbb{R}^d$ denote the feature vector of the i-th sample, where d is the number of software metrics after feature selection and $i = 1, 2, \dots, N$, with N representing the total number of samples. Let $f_{DNN}(\cdot)$, $f_{RF}(\cdot)$ and $f_{CNN}(\cdot)$ denote the trained FC-DNN, Random Forest and CNN models, respectively.

The FC-DNN estimates the probability that the i-th sample belongs to the faulty class as:

$$p_i^{DNN} = f_{DNN}(x_i), \qquad p_i^{DNN} \in [0,1], \tag{5}$$

where, p_i^{DNN} is the probability assigned by the FC-DNN to the faulty class.

The Random Forest predicts the probabilities of both classes as:

$$p_i^{RF} = [\, p_i^{(RF,0)}, p_i^{(RF,1)} \,] = f_{RF}(x_i), \tag{6}$$

where, $p_i^{(RF,0)}$ and $p_i^{(RF,1)}$ represent the predicted probabilities of the non-faulty and faulty classes, respectively, predicted by the Random Forest.

The outputs of the two base learners are concatenated to construct the meta-feature vector

$$z_i = [\, p_i^{DNN}, p_i^{RF} \,], \tag{7}$$

where, $z_i \in \mathbb{R}^3$ is the meta-feature vector, consisting of one probability generated by the FC-DNN and two class probabilities generated by the Random Forest and $[\cdot]$ denotes the vector concatenation operator.

The meta-feature vector is then provided as the input to the CNN meta-learner, which produces the final fault prediction according to o

$$\hat{y}_i = f_{CNN}(z_i), \qquad \hat{y}_i \in [0,1], \tag{8}$$

where, $\hat{y}_i$ denotes the final predicted probability that the i-th sample is faulty. A decision threshold of 0.5 is subsequently applied to assign the sample to either the faulty or non-faulty class.

Accordingly, for each software module, the FC-DNN predicts the probability of the faulty class, whereas the Random Forest estimates the probabilities of both classes. These three probability values are concatenated to form the meta-feature vector, which serves as the input to the CNN meta-learner. The CNN extracts higher-level interactions among these complementary predictions through its convolutional operation and produces the final software-fault prediction.

Algorithm 1 summarizes the complete workflow of the proposed FR-CNN stacking architecture. First,

the FC-DNN and Random Forest are independently trained using the same pre-processed training data. Their probability outputs are concatenated to construct the meta-feature representation, which is subsequently used to train the CNN meta-learner. During testing, the same sequence of operations is applied to unseen samples to generate the final fault prediction.

Algorithm 1. Pseudocode of the proposed FR-CNN stacking architecture

Input:
Preprocessed training set (X_{train}, Y_{train})
Preprocessed test set X_{test}

Output:
Predicted labels $\hat{Y}$

/ Train base learners */*
Train FC-DNN using (X_{train}, Y_{train})
Train Random Forest using (X_{train}, Y_{train})
/ Construct meta-features for training */*
$P_{\mathrm{DNN}_{train}} \leftarrow$ FC-DNN.predict (X_{train}) ▷ *Probability of the faulty class*
$P_{\mathrm{RF}_{train}} \leftarrow$ RF.predict_proba (X_{train}) ▷ *Probabilities of both classes*
$Z_{train} \leftarrow$ Concatenate $(P_{\mathrm{DNN}_{train}}, P_{\mathrm{RF}_{train}})$
/ Train meta-learner */*
Train CNN using (Z_{train}, Y_{train})
/ Predict on unseen instances */*
$P_{\mathrm{DNN}_{test}} \leftarrow$ FC-DNN.predict (X_{test})
$P_{\mathrm{RF}_{test}} \leftarrow$ RF.predict_proba (X_{test})
$Z_{test} \leftarrow$ Concatenate $(P_{\mathrm{DNN}_{test}}, P_{\mathrm{RF}_{test}})$
$\hat{Y} \leftarrow$ CNN.predict (Z_{test})
Return $\hat{Y}$

4. EVALUATION OF RESULTS

This Section presents a comprehensive evaluation of the proposed approach. It describes the dataset, software metrics, performance evaluation measures and implementation details. It also discusses the research questions, experimental results and corresponding findings.

4.1 Dataset

We have used the BugHunter, a new and reliable fault dataset collected by Ferenc R. et al. [28]. This dataset was created automatically using the OpenStaticAnalyzer tool and is publicly available. The BugHunter dataset assigns source code metrics to the code elements and it encompasses a larger amount of data compared to previous fault-prediction datasets. It includes a selection of 15 Java projects sourced from the GitHub repository, all of which are open source in nature. The projects of this dataset are the most common software, such as BroadleafCommerce, orientDB and Netty. BugHunter includes many instances at the method, class and file levels and contains four filters in addition to a basic and unfiltered case. Since the best-obtained results on this dataset were reported using the subtract filter, we selected the subtract-filter instances for six projects at the method level in our experiments and compared our results with the best-reported results in reference [28]. Every project in the BugHunter dataset includes 75 software metrics at the method level, which have been reduced after pre-processing step. Table 1 illustrates the dataset description for each project before and after the pre-processing. It also shows information about the quantity of software metrics, the imbalance ratio and the overall count of instances.

Table1. Description of the BugHunter projects before and after pre-processing (prep.).

Name of the project in the BugHunter	Total # of instances before prep.	# of faulty instances before prep.	# of non-faulty instances before prep.	Faulty ratio (%)	Imb. ratio	# of software metrics after prep.	Total # of instances after prep.	# of faulty instances after prep.	# of non-faulty instances after prep.
Netty	11171	2434	8737	21/78	3/58	57	16207	7470	8737
Neo4j	7030	1841	5189	26/18	3/14	57	9704	4515	5189
OrientDB	9445	2589	6856	27/41	2/64	59	12911	6055	6856
BroadleafCommerce	4709	1025	3684	21/76	3/59	59	6824	3140	3684
Ceylon-IDE-Eclipse	2087	508	1579	24/34	3/10	56	2972	1393	1579
MapDB	1456	480	976	32/96	2/03	57	1842	866	976

4.2 Software Metrics

The software metrics presented in the BugHunter dataset consist of static source-code metrics, code-smell metrics and code-clone metrics. Among 29 object-oriented metrics used at the method level of this dataset, for shortness, we listed some of them in Table 2. Table 3 presents the code-duplication metrics or basic clone-related metrics at the method level of the BugHunter.

4.3 Performance Evaluation Measures

To assess the results and compare the performance of the fault-prediction models, we employed measures including Accuracy, Precision, Recall and F1-score, Matthews Correlation Coefficient (MCC), Area Under the Receiver Operating Characteristic Curve (AUC-ROC), Area Under the Precision–Recall Curve (AUC-PR), Balanced Accuracy and G-Mean. Accuracy shows the proximity between an observed value and the true value, the proportion of the total count of correct predictions. Precision measures the degree of related results and recall measures how many related results are correct. F1-score represents a harmonic average between the precision and recall of a particular class. MCC provides a balanced evaluation by considering all four elements of the confusion matrix, making it particularly suitable for imbalanced datasets. AUC-ROC measures the model's ability to discriminate between defective and non-defective modules across different decision thresholds, while AUC-PR summarizes the trade-off between Precision and Recall. Balanced Accuracy computes the average of sensitivity and specificity and G-Mean evaluates the geometric mean of sensitivity and specificity, encouraging balanced-classification performance. These measures are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \tag{9}$$

$$\text{Precision} = \frac{TP}{TP + FP} \times 100 \tag{10}$$

$$\text{Recall} = \frac{TP}{TP + FN} \times 100 \tag{11}$$

$$\text{F1-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{12}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{13}$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{14}$$

where,

$$Sensitivity = \frac{TP}{TP + FN}, \qquad \text{Specificity} = \frac{TN}{TN + FP}$$

$$G - Mean = \sqrt{Sensitivity \times \text{Specificity}} \tag{15}$$

The AUC-ROC and AUC-PR metrics are threshold-independent measures computed from the Receiver

Operating Characteristic (ROC) curve and the Precision–Recall (PR) curve, respectively. Higher values indicate better discrimination capability, with a value of 1 representing perfect classification.

where TP is used to represent True Positives, FP signifies False Positives, FN denotes False Negatives and TN represents True Negatives.

In addition to the prediction performance metrics, we also evaluate the computational efficiency of the proposed framework using training time and testing time. Training time is defined as the total time required to fit a model using the training dataset, whereas testing time represents the total time required to generate predictions for the testing dataset. These measurements provide an indication of the computational cost and practical applicability of the evaluated models.

Table2. Object-oriented features in the Bughunter dataset [28].

Abbreviation	Full name
CLOC	Comment Lines of Code
LOC	Lines of Code
LLOC	Logical Lines of Code
NL	Nesting Level
NLE	Nesting Level Else-If
NII	Number of Incoming Invocations
NOI	Number of Outgoing Invocations
CD	Comment Density
DLOC	Documentation Lines of Code
TCD	Total Comment Density
NOS	Number of Statements

Table3. Basic clone-related features in the Bughunter dataset [28].

Abbreviation	Full name
CC	Clone Coverage
CCL	Clone Classes
CCO	Clone Complexity
CI	Clone Instances
CLC	Clone Line Coverage
CLLC	Clone Lines of Code
LDC	Lines of Duplicated Code
LLDC	Logical Lines of Duplicated Code

4.4 Implementation Details

This Sub-section pertains to the implementation details required for the experiments; we will specify the execution environment and set the parameter settings for both the presented model and the established baseline models. We initially divided each dataset into 80% training data and 20% independent test data during the pre-processing stage. We emphasize that the train-test split was performed before any data normalization or oversampling to prevent data leakage. Subsequently, during the model-training phase, 20% of the training data was reserved as a validation set to monitor model performance, tune hyper-parameters and prevent overfitting. Consequently, the effective data partition consisted of approximately 64% training, 16% validation and 20% test data. The validation set was derived exclusively from the training data and remained completely independent of the final test set, further minimizing any potential data leakage. For designing the FC-NN models, 3 to 6 layers were tested and evaluated. The Relu, Tanh and Swish activation functions were used for all layers except the last layer. For the output layer, however, the Sigmoid activation function was utilized. To train the FC-DNN, we used Adam, Adagrad and SGD optimizer functions. The learning rate employed for these optimizers was set to 0.001 and the binary cross entropy was adopted as the loss function. We tested the epochs from 1000 to 20000 to determine the appropriate epoch for network convergence. Different batch sizes from 10 to 14000 have been considered and different dropout rates have been investigated from 0.1 to 0.6 by increments of 0.1. In CNN, we tested the number of filters with values of 2, 4, 8, 15 and 32 and evaluated the length of each filter from 1 to 3. Additionally, we used the batch size values of 32, 64 and 128. The proposed method was implemented using TensorFlow 2.7, Keras 2.7 and Python 3.7. The Pandas, Numpy and Scikitlearn libraries were used in this work. In addition, the Matplotlib and Imblearn libraries were used for visualization and addressing the class-imbalance problem, respectively. Many experiments have been conducted using GPU and CPU. System-processor specifications include

an Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz.

For reproducibility, the CNN meta-learner of the proposed FR-CNN model employed a fixed architecture across all six projects to ensure a consistent stacking strategy. It consists of a one-dimensional convolutional layer with 8 filters (kernel size = 1) and ReLU activation, followed by a Flatten layer and a single-node output layer with a Sigmoid activation function. The model was trained using the Adam optimizer with binary cross-entropy loss for up to 100 epochs and a batch size of 64. Early stopping with a patience of 40 epochs based on the validation F1-score was applied to mitigate overfitting. To assess the effectiveness of the proposed model, we compared it with established ML and state-of-the-art baseline models (sub-sections 4.6.3 and 4.6.4). To ensure a fair comparison, all baseline models were implemented and evaluated under the same experimental conditions as the proposed method, including the same datasets, pre-processing pipeline, train-validation-test partitioning strategy, evaluation metrics and execution environment. The parameter settings for each baseline model are summarized below.

- The implementation of the traditional ML baseline models involved configuring them with default parameter values, utilizing the scikit-learn library.
- CBIL [54] includes an embedding layer with a dimension of 30, followed by a CNN layer with 20 filters, where each filter has a length of 10. Then, there is a max pooling layer, a BiLSTM layer with 32 LSTM units and a dense layer with the Sigmoid activation function. The tanh activation is used in both CNN and BiLSTM layers.
- The DL-based model [53] consists of 4 layers. The first layer has 20 nodes with tanh activation and the kernel initializer is uniform. The two hidden layers consist of 10 and 6 nodes, respectively, by using the relu activation. The output layer has a single node and the kernel initializer in the last three layers is normal.
- In 1D-CNN [18], 64 filters with a kernel size of 1 are chosen in each of two convolutional 1D layers. One max pooling layer with a pool size of 1 and two dense layers with activation ReLU and Sigmoid (last dense layer) were also used.
- The MLPs [50] consist of five layers, an input layer and three hidden layers employing the ReLU activation, as well as an output layer employing the Sigmoid activation. Following each of the first four layers, a dropout layer using the rate of 0.2 is included.

4.5 Research Questions

In evaluating the effectiveness of the presented model, the following research questions were formulated.

RQ1: Does tuning the FC-DNN and FR-CNN hyper-parameters increase the performance?

RQ2: Does proposed FR-CNN hybrid model achieve better performance compared to the individual learners?

RQ3: How effective is the proposed model in improving the prediction performance compared to the ML baseline models and the existing performance on the BugHunter dataset?

RQ4: Does proposed FR-CNN achieve better performance compared to the state-of-the-art models across all given projects in the BugHunter dataset?

RQ5: How does the proposed model compare with traditional machine-learning and state-of-the-art deep-learning approaches in terms of prediction performance, computational efficiency and model complexity on the BugHunter dataset?

RQ6: Does the proposed FR-CNN provide statistically significant improvements over existing software fault-prediction approaches?

RQ7: Does the proposed FR-CNN achieve an effective balance between prediction performance and computational efficiency while demonstrating the contribution of each architectural component?

4.6 Results and Discussion

The results of the proposed FC-DNN and FR-CNN models, along with the answers to RQ1 - RQ4, have been presented and discussed in the following sub-sections.

4.6.1 RQ1: Effect of Tuning the Hyper-parameters

To address the first research question, we present the experiments in subsections A and B.

A. Hyper-parameter Tuning Process for FC-DNN

To systematically evaluate the impact of hyper-parameter tuning on the performance of the proposed FC-DNN, we employed a comprehensive grid search strategy across a wide range of hyper-parameter configurations. The tuning process was conducted independently for each of the six projects.

Search Strategy and Parameter Ranges: We began with a base architecture and iteratively manipulated individual hyper-parameters while keeping others constant. Once the optimal value for a given hyper-parameter was identified, it was fixed and the tuning cycle was repeated for the next hyper-parameter. This iterative refinement process ensured that each hyper-parameter converged to its optimal value. The parameter ranges explored were as follows: number of epochs from 1,000 to 20,000 (with increments of 500-1000); batch size from 10 to 14,000 (with smaller increments for smaller batch sizes and larger steps for larger batch sizes); dropout rate from 0.1 to 0.6 (in increments of 0.1); number of hidden layers from 3 to 6; activation functions including ReLU, Tanh and Swish for hidden layers (with Sigmoid fixed for the output layer); and optimizer functions including Adam, Adagrad and SGD. This exhaustive exploration resulted in over 200 distinct experimental configurations being evaluated across all projects.

Validation Procedure and Selection Criteria: During the hyper-parameter tuning phase, a fixed validation split of 20% from the training data was employed to monitor training progress and prevent overfitting. For each hyperparameter configuration, the model was trained on the training sub-set and evaluated on the validation sub-set, with Accuracy and F1-score considered as the primary evaluation metrics. The final selected model was subsequently evaluated once on the independent test set. Table 4 summarizes the performance of the FC-DNN models obtained using the optimal hyper-parameter configurations for each project in terms of Accuracy, Precision, Recall and F1-score. These results demonstrate that different hyper-parameter configurations yield different prediction performances for the proposed FC-DNN model across the six BugHunter projects.

B. Hyper-parameter Tuning for FR-CNN Meta-learner

For the FR-CNN hybrid model, we investigated the impact of three key hyper-parameters of the CNN meta-learner: the number of filters (ranging from 2 to 32), kernel size (ranging from 1 to 3) and batch size (ranging from 32 to 128). Similar to the FC-DNN tuning process, we evaluated each configuration using the same data split and performance metrics. The tuning was performed by modifying one hyper-parameter at a time while keeping others fixed until reaching the highest performance. The average performance in terms of accuracy and F1-score was calculated and visualized on bar graphs (Figures 5-7). The results indicate that variations in these parameters do not yield significant performance enhancement, suggesting that the FR-CNN is relatively robust to these hyper-parameter choices.

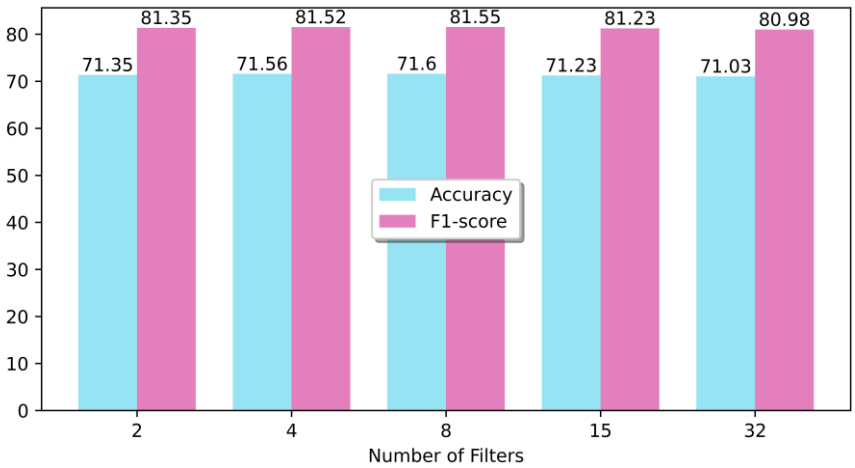


Figure 5. Effect of the number of CNN filters on the average Accuracy and F1-score of the proposed FR-CNN.

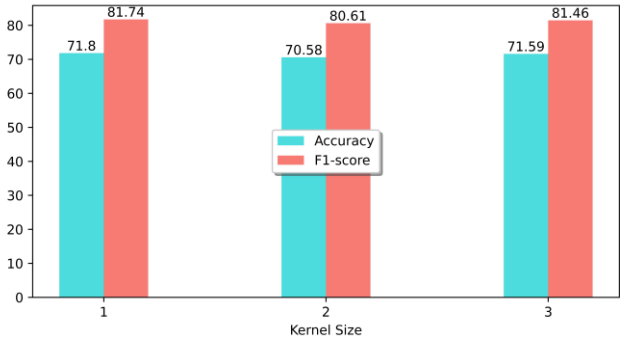


Figure 6. Effect of CNN kernel size on the average Accuracy and F1-score of the proposed FR-CNN.

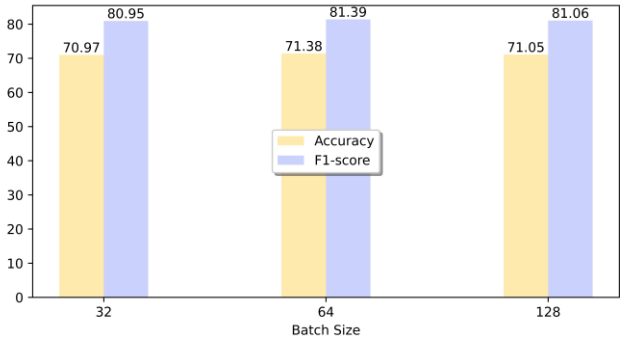


Figure 7. Effect of CNN batch size on the average Accuracy and F1-score of the proposed FR-CNN.

4.6.2 RQ2: Performance Comparison between the Presented FR-CNN and Individual Learners

In Tables 5 and 6, we present the results of the optimal FC-DNN model from the previous sub-section, the FR-CNN hybrid model and the RF model on six projects. The highest values of the average accuracy and F1-score are highlighted in bold. By comparison, the proposed FR-CNN hybrid model outperformed the proposed FC-DNN on the Neo4j orientDB, BroadleafCommerce, Ceylon-IDE-Eclipse and Netty projects, demonstrating improvements of up to 4.78% in accuracy (from 65.31 to 70.09) and up to 4.04% in F1-score (77.44 to 81.48). Compared to the RF, the FR-CNN improved the prediction performance on all six projects, especially on Ceylon-IDE-Eclipse. The FR-CNN demonstrated an improvement of up to 5.26% in accuracy (from 64.83% to 70.09%) and enhanced the F1-score by up to 4.35% (from 77.13% to 81.48%) on RF. We can conclude that the FR-CNN is a robust model that can enhance the prediction accuracy of any individual learner participating in the model, in terms of accuracy and F1-score, by covering their weaknesses. Compared to the other projects, the presented FC-DNN model outperformed the presented FR-CNN model on MapDB, exhibiting a 4.45% increase in accuracy (from 60.61 to 65.06) and a 5.3% improvement in F1-score (from 72.42 to 77.72). This behavior can be attributed to the hierarchical stacking mechanism adopted by the proposed FR-CNN. The CNN-based meta-learner combines the outputs of both the FC-DNN and the RF; therefore, its effectiveness depends on the complementary predictive information provided by the two base learners. The feature importance analysis reveals that RF and FC-DNN focus on different aspects of the feature space, confirming that their complementary strengths justify their use as base learners in the hierarchical FR-CNN architecture. However, on the MapDB dataset, the RF exhibits comparatively weaker predictive performance than the FC-DNN, limiting the amount of complementary information available to the meta-learner. As a result, the hybrid model is unable to further improve upon the stronger standalone FC-DNN on this particular dataset. This observation does not indicate a limitation of the stacking strategy itself; rather, it highlights that the effectiveness of hierarchical ensembles depends on the quality and diversity of their constituent base learners. When one base learner already captures most of the discriminative information while the other contributes relatively weaker predictions, the potential performance gain of the ensemble naturally becomes limited. Our analysis further reveals that the important features identified by the FC-DNN on the MapDB dataset are more discriminative than those selected by the RF, enabling the FC-DNN to capture the underlying fault patterns more effectively on this project. In contrast, on the remaining BugHunter projects, the complementary strengths of the two base learners enable the proposed FR-CNN to outperform the individual models. This interpretation is further supported by the

feature-importance analysis presented in Figures 8–10. The FC-DNN consistently identifies more discriminative features with lower uncertainty than the RF on the MapDB project, indicating that the neural network captures the underlying characteristics of this relatively small dataset more effectively. Consequently, the additional information contributed by the RF is insufficient to further improve the ensemble prediction on this project, whereas on the remaining BugHunter projects the complementary strengths of the two base learners enable FR-CNN to outperform the individual models.

4.6.3 RQ3: Performance Comparison of the Presented Model and the Traditional Baseline Models

To answer the third research question, we conduct the performance comparison between the proposed model and the seven established baseline classification models: KNN, NB, DT, LR, RF, GA and SVM. We also evaluate the performance of our model against the best performance achieved by [28] using 11 traditional ML algorithms, including RandomForest, DecisionTable, J48 (C4.5), RandomTree and SimpleLogistic and present the results in Tables 5 and 6. Additionally, we conduct a comparison between the best baseline model when using the scikit-learn library, which we used and the best ML algorithm when using the Weka library, as utilized by [28].

On average (Tables 5 and 6), the presented model enhanced the accuracy and F1-score of seven classification baseline models by 12.24% and 15.1%, respectively. In terms of F1-score (Table 6), the proposed model outperformed the best traditional ML algorithms [28] by 20.09%. By comparison, our proposed model increased the performance value from 53.95, using the RandomTree algorithm [28], to 81.48 on the Ceylon-IDE-Eclipse project, improving the performance by 27.53%. Also, the best baseline model used in this study, the RF, improved the average F1-score of the best traditional ML algorithms used in [28] by 18.28% on the selected projects.

4.6.4 RQ4: The Performance of the Presented Model against State-of-the-Art Approaches

To answer the fourth research question, we compare the proposed FR-CNN with five recent state-of-the-art software fault-prediction approaches: Hybrid model of Convolutional neural network and Bidirectional Long short-term memory (CBIL) [54], Deep Learning-based model (DL-based model) [53], 1-Dimensional Convolutional Neural Network (1D-CNN) [18] and Multi-layer perceptron's (MLPs) [50] and SEA-based metrics [30]. The comparative results are presented in Tables 7 and 8. Overall, the proposed FR-CNN demonstrates competitive prediction performance across the six BugHunter projects. Although no single method consistently achieves the highest performance on every project, FR-CNN provides a favorable balance between predictive effectiveness and computational efficiency, as further supported by the complexity analysis presented in the following sub-section.

Compared with the CBIL model, FR-CNN achieves competitive prediction performance across the evaluated BugHunter projects. In particular, it outperforms CBIL on the BroadleafCommerce project, achieving improvements of up to 8.17% (from 77.07 to 85.24) in accuracy and 4.53% (from 86.36 to 90.89) in F1-score, whereas CBIL achieves higher F1-scores on several of the remaining projects. The average performance of the proposed model closely approximates that of the CBIL model. Surprisingly, in the rest of the projects, the CBIL outperforms our proposed approach, which can be related to the specific structure of the CBIL model, comprising a combination of CNN and BiLSTM layers, resulting in varying performance across different projects. Nevertheless, it is important to highlight that, we took measures to prevent overfitting and minimize error during the training phase of our model to enhance its performance. Figure 11 presents the comparison of the model loss between our proposed approach and CBIL across BroadleafCommerce dataset, with the left graph corresponding to our proposed approach and the right graph corresponding to CBIL. Graphs show the same trend and fluctuation for other projects. Compared with the published optimization behavior of CBIL, the training and validation losses of which exhibit more noticeable oscillations across the BugHunter projects, the proposed FR-CNN consistently converges more smoothly. This stable optimization behavior demonstrates that the proposed hierarchical architecture achieves competitive prediction performance without sacrificing training stability or exhibiting signs of overfitting. We can conclude that our proposed approach achieved the lowest training and evaluation errors across various datasets, indicating excellent performance during the training phase without any overfitting issues. We included dropout layers in the proposed FC-DNN layers and used an early stopping process during training the FR-CNN model to ensure this. However, the prediction performance of the proposed approach needs improvement in the rest of the projects and we will consider it for future work.

Compared to the DL-based model, our proposed approach successfully improved the prediction performance in five projects: Ceylon-IDE-Eclipse, BroadleafCommerce, Netty, MapDB and OrientDB. The most significant increase in prediction performance was observed in the Ceylon-IDE-Eclipse project, with a 9.57% increase (from 60.52 to 70.09) in accuracy and a 7.15% increase (from 74.33 to 81.48) in F1-score. This was followed by BroadleafCommerce, which exhibited a 9.13% increase (from 76.11 to 85.24) in accuracy and a 6.36% increase (from 84.53 to 90.89) in F1-score.

Compared to the 1D-CNN, the proposed method improves the prediction performance in terms of accuracy and F1-score on the BroadleafCommerce, Ceylon-IDE-Eclipse and Netty projects. Our proposed method has better accuracy on the BroadleafCommerce and Netty, by 9.87% (from 75.37 to 85.24) and 8.59% (from 67.74 to 76.33), respectively. In terms of F1-score, the proposed method again shows superior performance on the Netty and BroadleafCommerce by 6.8% (from 78.73 to 85.53) and 6.76% (from 84.13 to 90.89), respectively. These findings indicate that the presented model is more efficient and effective than 1D-CNN, in terms of classification accuracy and F1-score on the aforementioned projects.

Compared to MLPs, the presented model improved the average accuracy and f1-score by 7.43% and 6.61%, respectively, on the Neo4j, Ceylon-IDE-Eclipse and MapDB projects.

Compared to SEA-based metrics, the proposed method improves the prediction performance in five projects: Neo4j, BroadleafCommerce, Ceylon-IDE-Eclipse, Netty and MapDB. In terms of F1-score, the proposed method shows superior performance on the Ceylon-IDE-Eclipse and BroadleafCommerce by 19.95% (from 61.53 to 81.48) and 15.77% (from 75.12 to 90.89), respectively.

4.6.5 Additional Performance Evaluation and Computational Analysis

To answer the fifth research question, we evaluate the effectiveness of the proposed approach and state-of-the-art approaches against traditional ML algorithms on the selected projects. In terms of accuracy, the presented model and state-of-the-art models demonstrate superiority against the ML baseline models with 69.09%. In terms of F1-score, they again outperformed ML models with 78.96%. On average, the proposed model and state-of-the-art models improved the accuracy and F1-score of the traditional ML baseline algorithms by 9.26% and 12.75%, respectively.

It is also worth clarifying the comparatively low average MCC values reported in Figure 12 relative to Balanced Accuracy, AUC-ROC and G-Mean. Unlike Balanced Accuracy, which is bounded below by 0.5 for a random classifier and therefore tends to remain in a moderate range (0.5–0.6) even for weak models, MCC is a chance-corrected measure in which a value of 0 indicates performance statistically indistinguishable from random guessing and it penalizes any imbalance among the four confusion-matrix cells multiplicatively rather than additively. Consequently, on projects where a classifier's predictions are only marginally better than chance for the minority (faulty) class, Balanced Accuracy can still appear moderately above 0.5, while the corresponding MCC collapses toward zero. This effect is most evident on the two smallest projects in our study, Ceylon-IDE-Eclipse and MapDB, where FR-CNN obtains an MCC of only 0.051 and 0.040, respectively (Balanced Accuracy of 0.520 and 0.519), compared with an MCC of 0.496 (Balanced Accuracy of 0.731) on BroadleafCommerce. Notably, near-zero or even negative MCC values on Ceylon-IDE-Eclipse are observed across nearly all compared models—not only FR-CNN—including NB (−0.062), SVM (−0.053), KNN (−0.050), 1D-CNN (−0.036) and CBIL (−0.043), indicating that this is a dataset-level property related to the limited size and difficulty of these two projects rather than a weakness specific to the proposed model. Since Figure 12 reports a simple (unweighted) average across the six BugHunter projects, the near-zero MCC values on these two smaller, harder projects disproportionately pull down the six-project average MCC relative to the other metrics, which are individually less sensitive to this near-chance behavior. The per-project MCC and Balanced Accuracy values underlying this analysis were verified by the authors against the raw confusion-matrix outputs for all six projects and all twelve models.

The practical applicability of the proposed framework is further demonstrated through computational complexity and execution-time analyses, as presented in Figs. 13 and 14. Compared with recent deep-learning approaches, FR-CNN requires substantially fewer trainable parameters and shorter training time while maintaining a testing time below one second on average. The proposed model employs only 41 trainable parameters in its CNN meta-learner, whereas 1D-CNN, CBIL and the DL-based model require approximately 22.7K, 236K and 1.5K trainable parameters, respectively. It should be noted that

these parameters correspond only to the lightweight CNN meta-learner, since the FC-DNN base learners are pre-trained offline using their optimized hyper-parameters and subsequently reused with fixed weights during deployment. Consequently, practitioners are not required to retrain the FC-DNN components, reducing computational overhead while preserving the representational capability of deep learning. Overall, the complexity analysis demonstrates that FR-CNN achieves competitive prediction performance while requiring substantially fewer trainable parameters and shorter training time than the recent deep-learning-based approaches evaluated in this study. These results indicate that the proposed hierarchical framework provides a favorable balance between predictive performance and computational efficiency, making it suitable for practical software-engineering environments where computational resources, training time and model complexity are important considerations.

4.6.6 Statistical Significance Analysis

To verify that the performance improvements of FR-CNN are statistically meaningful rather than due to random variation, we conducted non-parametric significance testing across the six BugHunter projects. Results were presented in Table 9. First, we applied the Friedman test, which evaluates whether performance differences exist among all compared models across the six projects. The test yielded a highly significant result for both accuracy ($\chi^2 = 40.31$, $p < 0.00001$) and F1-score ($\chi^2 = 41.42$, $p < 0.00001$), confirming that the observed differences among models are unlikely to have occurred by chance. We then conducted pairwise Wilcoxon signed-rank tests comparing FR-CNN against each baseline model (KNN, NB, DT, LR, RF, GA, SVM) across the six projects, applying the Benjamini-Hochberg false-discovery rate (FDR) correction to account for multiple comparisons. FR-CNN significantly outperformed all seven traditional ML baselines on both accuracy and F1-score (adjusted $p = 0.036$ for all comparisons), indicating consistent superiority across projects rather than dataset-specific gains. The comparison between FR-CNN and the standalone FC-DNN did not reach statistical significance (accuracy: $p = 0.31$; F1-score: $p = 0.44$), consistent with the MapDB result discussed in Sub-section 4.6.2, where the FC-DNN outperformed the hybrid model due to the comparatively weaker RF performance on that project. We note that, with six paired projects, the Wilcoxon test has a minimum achievable two-sided p-value of 0.031; while this constrains the granularity of the reported significance levels, the consistent direction of the effect across all projects, combined with the significant Friedman omnibus result, supports the robustness of the observed improvement.

4.6.7 Ablation Study: Meta-learner Comparison and Component Contribution

To directly assess the design choices underlying the proposed FR-CNN architecture, we conducted two complementary ablation studies across all six BugHunter projects.

A. Meta-learner Comparison: Holding the base learners and the meta-feature vector $z_i = [p_{DNN}, p_{RF,0}, p_{RF,1}]$ fixed, we replaced the CNN meta-learner with three alternatives: a feed-forward DNN of matched capacity (Dense(8, ReLU) → Dense(1, Sigmoid)), a Logistic Regression classifier and a parameter-free unweighted average of p_{DNN} and $p_{RF,1}$. The CNN and DNN variants, being stochastic, were each trained over five random seeds; we report the mean, with the standard deviation across seeds not exceeding 0.49 F1 points for any project, confirming that the training procedure converges reliably. Table 10 reports the per-project and average F1-score, together with the number of trainable parameters and training time for each meta-learner.

The results show that the three learnable meta-learners (CNN, DNN, Logistic Regression) perform comparably, with average F1-scores within 0.3 points of one another (80.74–81.00%) and on average outperform simple unweighted averaging by approximately one point of F1-score and accuracy across the six projects. This margin is not uniform across every project: on MapDB and to a lesser extent OrientDB, simple averaging performs on par with or slightly better than the learnable meta-learners, whereas on the remaining four projects the learnable meta-learners provide a clearer advantage of 1.4–2.4 F1 points. This indicates that the principal benefit of the meta-learning stage comes from learning a combination of the base-learner outputs rather than from the specific functional form of the meta-learner. We retain the CNN meta-learner in the proposed FR-CNN, because it matches or exceeds the simpler learnable alternatives while requiring only 41 trainable parameters and a training time under 34 seconds on average (Table 10), consistent with the lightweight-inference goal already demonstrated in Sub-section 4.6.5 (Figs. 13–14). We note that Logistic Regression achieves a marginally higher average F1-score (81.00% *vs.* 80.83%) at a fraction of the parameter and training cost; we report this transparently,

as the meaningful architectural conclusion is that a learned combination rule is beneficial, while the specific choice among CNN, DNN and Logistic Regression has a comparatively minor effect on predictive performance.

B. Component Contribution: To isolate the contribution of each element of the hierarchical architecture, we compared FC-DNN alone, RF alone and the FC-DNN+RF combination under three different combiners: simple averaging, a Logistic Regression meta-learner and the proposed CNN meta-learner. Table 11 reports the per-project and average F1-score for each configuration.

Combining FC-DNN and RF, regardless of the combiner used, improves over FC-DNN alone on five of the six projects, with gains of up to 4.7 F1 points on Netty and matches or slightly exceeds RF alone on all six projects. The only exception is MapDB, where every combined configuration performs worse than the standalone FC-DNN. This is consistent with and independently corroborates our analysis in Sub-section 4.6.2, which attributes this behavior to RF's comparatively weaker predictive performance on the MapDB project, which limits the complementary information available to any meta-learner regardless of its architecture. Taken together, Tables 10 and 11 confirm that (i) each base learner contributes meaningfully to the hybrid architecture except where one base learner is markedly weaker than the other and (ii) the specific meta-learner used to combine them is a secondary, though non-negligible, design choice.

5. Threats to Validity

To maintain scientific rigor, it is important to acknowledge potential limitations across four key areas: internal, external, construct and conclusion validity.

5.1 Internal Validity

A primary internal threat is the data pre-processing pipeline. We removed nominal features, such as Hash and LongName, and also filtered out strictly constant features (i.e., those with no variance across all instances). While this is a standard and necessary step, it assumes that any feature with zero variance is entirely non-informative. This assumption is reasonable, but simplistic, as such features could theoretically have predictive value in rare cases.

5.2 External Validity

External validity concerns the generalizability of our results. Our experiments were conducted on six selected projects from the BugHunter dataset at the method level. Furthermore, the results may not generalize to other widely used datasets, such as PROMISE or NASA, which have different characteristics and features. The choice of granularity level (method, class or file) could also influence the outcomes, as fault patterns may vary across these levels.

5.3 Construct Validity

The primary construct we aim to predict is "fault-proneness," which we operationalize as a binary outcome (i.e., whether a module contains bugs or not). This simplification is common in the field, but does not distinguish between the severity or count of bugs. Additionally, the selected software metrics (such as coupling, cohesion and code smells) are indirect indicators of complexity and maintainability and may not fully capture all factors contributing to fault-proneness.

5.4 Conclusion Validity

We mitigated conclusion-validity threats by evaluating our models using multiple metrics (Accuracy, Precision, Recall, F1-score and MCC) rather than relying on a single measure. To ensure reproducibility and reduce variability, a fixed random seed (42) was used for all stochastic processes. Furthermore, an early stopping mechanism was applied during training to prevent overfitting. The use of training/validation loss curves further supports the stability of our model's training process. While the FR-CNN model shows competitive performance, it does not always consistently outperform all baseline models and we have reported these comparisons transparently in the Results Section.

We also note a limitation regarding the stability of the headline improvement figures reported in this study (e.g., the 19.95% F1-score improvement over SEA-based metrics on Ceylon-IDE-Eclipse, Sub-

section 4.6.4 and the comparisons against the traditional ML and state-of-the-art baselines in Sub-sections 4.6.3 and 4.6.4). These headline results, like the fixed-seed setting described above, are reported from a single run per model (seed = 42) rather than averaged across multiple random seeds, since several of the compared baseline results are literature-reported values for which repeated-seed statistics are not available. We therefore treat these point-estimate improvements as indicative rather than as precise, variance-quantified effect sizes. Where repeated-seed evaluation was feasible — i.e., for the FR-CNN and DNN meta-learners in the ablation study of Sub-section 4.6.7 — Table 10 reports the mean F1-score across five random seeds for each project and we additionally state in the accompanying text of Sub-section 4.6.7 that the observed standard deviation across seeds did not exceed 0.49 F1 points for any project, suggesting that the training procedure itself is stable, even though the per-project standard deviations themselves are not tabulated. We nonetheless flag the single-seed nature of the remaining headline comparisons as a limitation and recommend that future replications average results over multiple seeds, particularly for the comparisons against externally reported baselines.

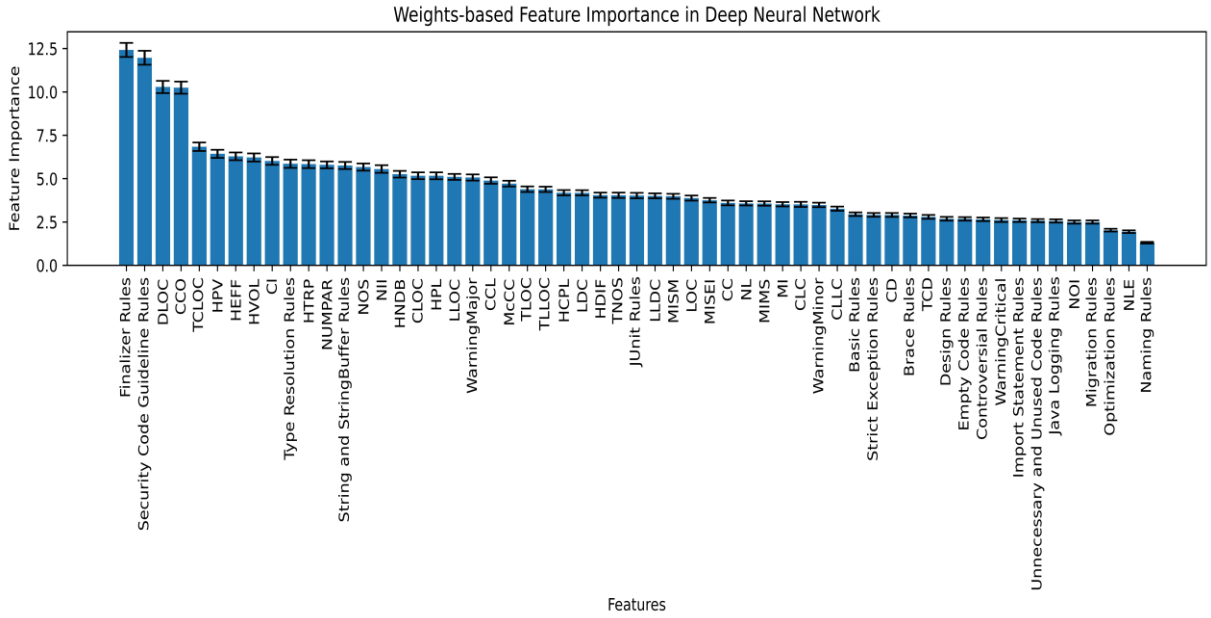


Figure 8. Comparison of the feature importance identified by the FC-DNN, illustrating its complementary feature representations in the proposed FR-CNN. Error bars represent the standard deviation of input layer weights, indicating uncertainty in the feature importance.

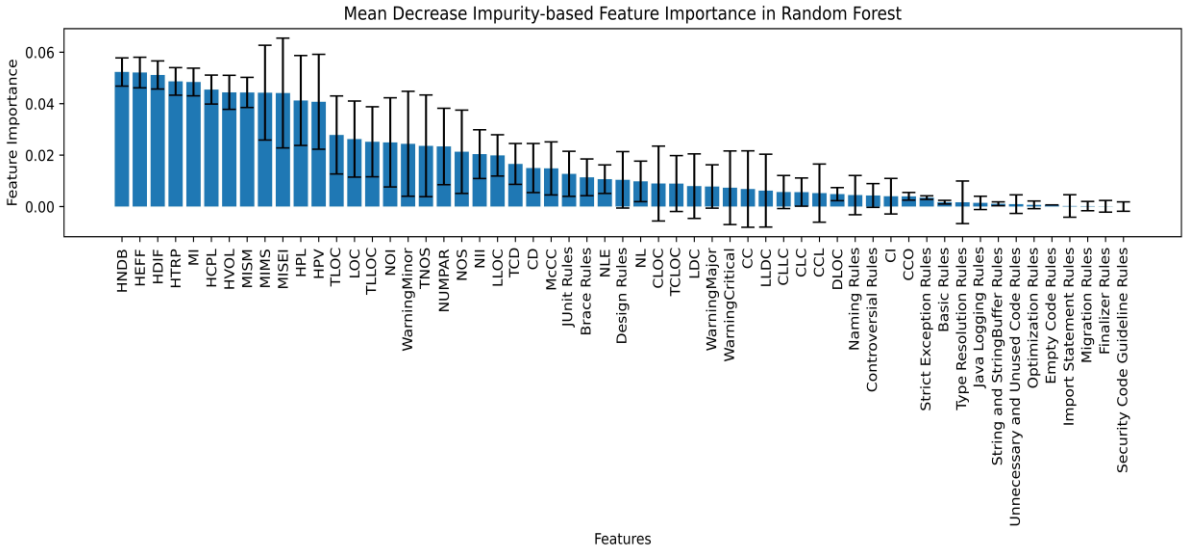


Figure 9. Comparison of the feature importance identified by the RF, illustrating its complementary feature representations in the proposed FR-CNN. Error bars represent the standard deviation of feature-importance scores obtained from individual trees within the RF ensemble, indicating uncertainty in the feature importance.

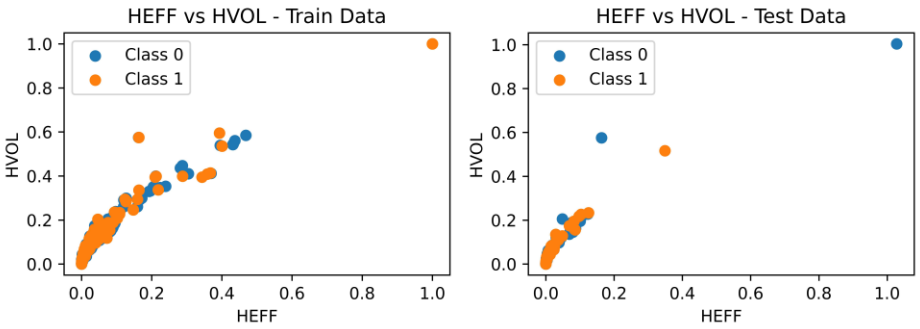


Figure 10. Faulty or non-faulty class association with HEFF and HVOL which are common among the 10 most important features shared with both the RF and FC-DNN.

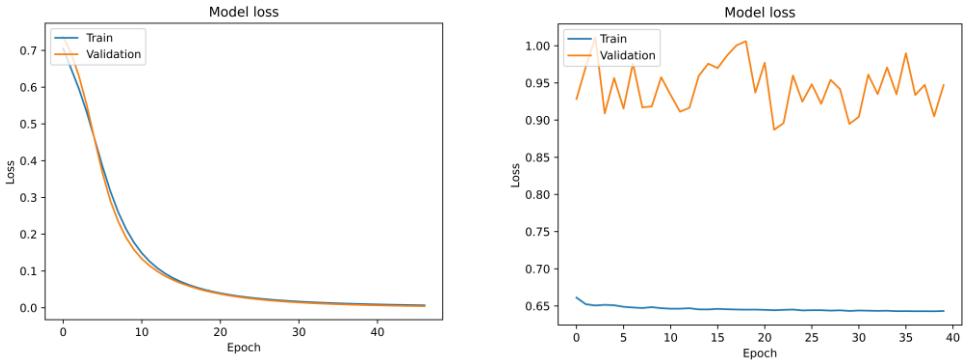


Figure 11. Training and validation loss curves of a) the proposed FR-CNN model and b) CBIL state-of-the-art model, illustrating the convergence behavior during training.

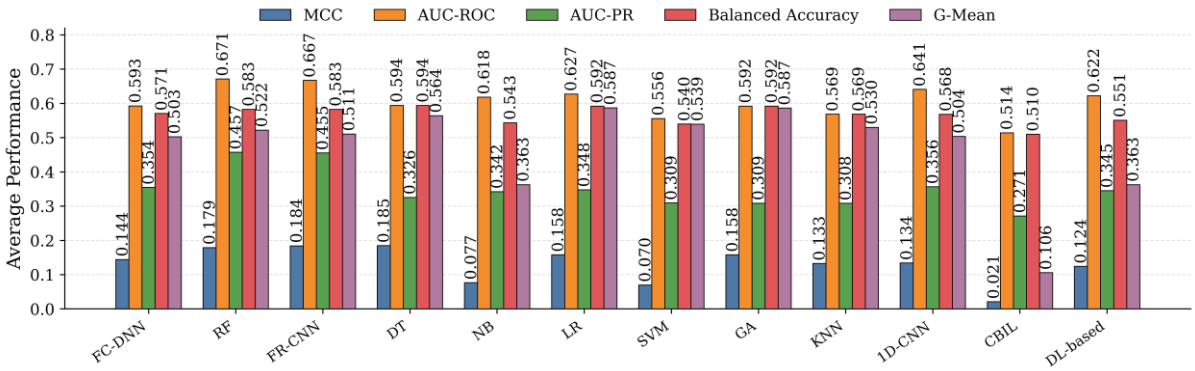


Figure 12. Comparison of the average performance of the proposed FR-CNN and baseline models using additional evaluation metrics (MCC, AUC-ROC, AUC-PR, Balanced Accuracy and G-Mean) across the six selected BugHunter projects.

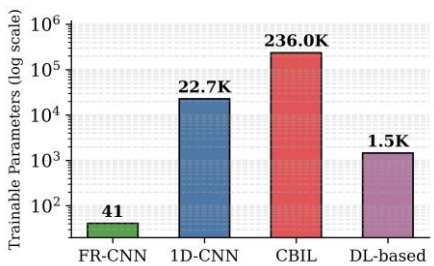


Figure 13. Comparison of the trainable parameters of the proposed FR-CNN and recent deep-learning approaches. The reported parameters for FR-CNN correspond only to the lightweight CNN meta-learner, while the pre-trained FC-DNN base learners are reused with fixed weights during deployment.

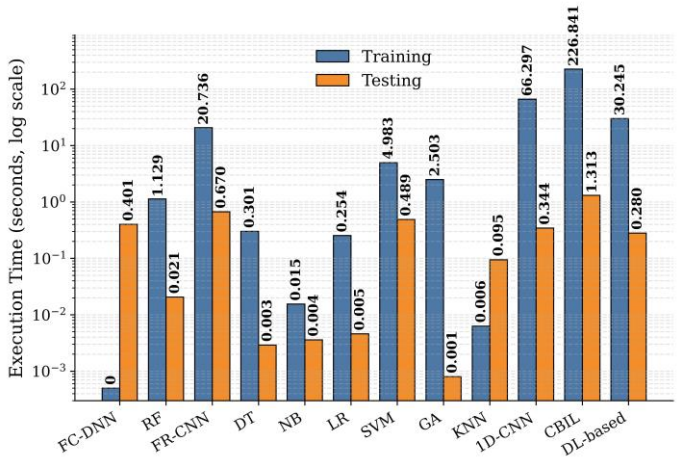


Figure 14. Comparison of the average training time and testing time of the proposed FR-CNN, traditional machine-learning algorithms and recent deep-learning approaches across the six selected BugHunter projects.

Table 4. Optimal hyper-parameter for each project achieved by more than 200 experiments during FC-DNN model tuning.

Project name	#epochs	Batch size	Dropout rate	optimizer	#layers	Activation function	Acc.	Prec.	Rec.	F1.
BroadleafCommerce	9500	2000	0.1	Adam	4	Relu	81.52	88.15	88.39	88.27
Ceylon-IDE-Eclipse	5000	2600	0.4	Adam	3	Relu	65.31	82.45	73.02	77.44
Neo4j	8000	8500	0.2	Adam	3	Relu	67.70	81.15	76.58	78.80
OrientDB	7500	1000	0.6	Adam	3	Swish	67.33	81.56	74.82	78.05
Netty	5000	7000	0.1	Adam	3	Swish	71.90	82.35	81.88	82.11
MapDB	2000	1000	0.6	Adam	3	Tanh	65.06	88.55	69.26	77.72

Table 5. Accuracy (%) of the proposed FC-DNN and FR-CNN models *vs.* traditional ML baseline models.

Name of project	FC-DNN	FR-CNN	KNN	NB	DT	LR	RF	GA	SVM
Neo4j	67.70	71.76	61.73	65.29	67.14	61.37	69.77	61.02	60.31
OrientDB	67.33	68.39	61.14	45.47	66.11	57.54	66.59	57.64	58.28
BroadleafCommerce	81.52	85.24	73.24	21.76	81.42	64.86	84.71	64.75	66.34
Ceylon-IDE-Eclipse	65.31	70.09	54.54	28.46	61.96	55.02	64.83	54.30	55.26
Netty	71.90	76.33	64.42	60.40	72.08	58.34	75.21	59.59	57.76
MapDB	65.06	60.61	52.05	33.21	58.56	57.87	59.24	57.53	56.16
Average	**69.80**	**72.07**	61.18	42.43	67.87	59.16	**70.05**	59.13	59.01

Table 6. F1-score (%) of the proposed FC-DNN and FR-CNN models *vs.* traditional ML baseline models.

Name of project	FC-DNN	FR-CNN	[28]	KNN	NB	DT	LR	RF	GA	SVM
Neo4j	78.80	82.06	60.86	71.71	74.92	77.59	69.98	80.35	69.65	68.82
OrientDB	78.05	79.23	62.36	70.56	41.60	76.41	64.48	77.58	64.66	65.13
Broadleaf-Commerce	88.27	90.89	73.66	81.60	2.38	88.10	73.45	90.47	73.13	75.93
Ceylon-IDE-Eclipse	77.44	81.48	53.95	66.19	6.26	74.06	66.06	77.13	66.07	64.38
Netty	82.11	85.53	64.12	75.17	71.21	81.93	68.19	84.72	69.40	67.15
MapDB	77.72	72.42	56.10	60.89	9.30	69.21	68.21	70.47	67.19	65.59
Average	**80.39**	**81.93**	61.84	71.02	34.27	77.88	68.39	**80.12**	68.35	67.83

Table 7. Accuracy (%) of the proposed FC-DNN and FR-CNN models *vs.* state-of-the-art approaches.

Name of project	FC-DNN	FR-CNN	CBIL [54]	DL-based model [53]	1D-CNN [18]	MLPs [50]
Neo4j	67.70	71.76	73.54	72.54	73.32	63.39
OrientDB	67.33	68.39	71.04	67.28	71.20	-
BroadleafCommerce	**81.52**	**85.24**	77.07	76.11	75.37	-
Ceylon-IDE-Eclipse	65.31	70.09	72.00	60.52	61.96	62.67
Netty	71.90	76.33	77.98	71.90	67.74	-
MapDB	65.06	60.61	68.15	62.32	67.12	58.56
Average	69.80	**72.07**	**73.29**	68.44	69.45	61.54

Table 8. F1-score (%) of the proposed FC-DNN and FR-CNN models *vs.* state-of-the-art approaches.

Name of project	FC-DNN	FR-CNN	CBIL [54]	DL-based model [53]	1D-CNN [18]	MLPs [50]	SEA-based metrics
Neo4j	78.80	82.06	84.61	82.43	84.38	75.36	73.84
OrientDB	78.05	79.23	83.03	77.54	83.17	-	82.30
BroadleafCommerce	**88.27**	**90.89**	86.36	84.53	84.13	-	75.12
Ceylon-IDE-Eclipse	77.44	81.48	83.72	74.33	75.11	74.83	61.53
Netty	82.11	85.53	87.63	82.57	78.73	-	73.58
MapDB	77.72	72.42	81.05	74.88	78.85	71.25	66.41
Average	80.39	**81.93**	**84.4**	79.38	80.72	73.81	72.13

Table 9. Statistical-significance testing results comparing FR-CNN against baseline models across all six BugHunter projects.

Comparison (FR-CNN *vs.*)	**Accuracy p-value**	**Accuracy (FDR-adj.)**	**F1-score p-value**	**F1-score (FDR-adj.)**	**Significant?**
FC-DNN	0.3125	0.3125	0.4375	0.4375	No
KNN	0.0312	0.0357	0.0312	0.0357	Yes
Naive Bayes	0.0312	0.0357	0.0312	0.0357	Yes
Decision Tree	0.0312	0.0357	0.0312	0.0357	Yes
Logistic Regression	0.0312	0.0357	0.0312	0.0357	Yes
Random Forest	0.0312	0.0357	0.0312	0.0357	Yes
Genetic Algorithm	0.0312	0.0357	0.0312	0.0357	Yes
SVM	0.0312	0.0357	0.0312	0.0357	Yes

Table 10. F1-score (%) of the proposed CNN meta-learner *versus* simpler alternatives, using the identical base-learner outputs across the six BugHunter projects.

Project	**CNN (proposed)**	**DNN**	**Logistic Regression**	**Simple averaging**
BroadleafCommerce	90.33	90.20	90.28	88.56
Ceylon-IDE-Eclipse	79.88	79.88	79.70	77.45
MapDB	70.39	69.97	70.10	73.21
Neo4j	81.29	81.24	81.53	79.89
Netty	85.02	85.04	85.68	83.21
OrientDB	78.05	78.13	78.74	78.18
Average	**80.83**	80.74	81.00	80.08
Trainable parameters	41	41	4	0
Average training time (s)	33.63	33.16	0.03	0.00

Table 11. F1-score (%) contribution of each component of the hierarchical architecture across the six BugHunter projects.

Project	FC-DNN alone	RF alone	FC-DNN+RF, averaging	FC-DNN+RF, LR meta-learner	FC-DNN+RF, CNN meta-learner (FR-CNN)
BroadleafCommerce	87.61	90.08	88.56	90.28	90.22
Ceylon-IDE-Eclipse	76.28	78.53	77.45	79.70	79.76
MapDB	77.38	69.80	73.21	70.10	70.10
Neo4j	78.49	80.11	79.89	81.53	81.30
Netty	80.34	84.83	83.21	85.68	85.02
OrientDB	77.39	77.71	78.18	78.74	78.10
Average	79.58	80.18	80.08	81.00	80.75

6. Conclusion and Future Work

In this study, we proposed FR-CNN, a hierarchical hybrid framework for method-level software-fault prediction on the BugHunter dataset. The framework integrates a fine-tuned FC-DNN, Random Forest and a CNN-based meta-learner following comprehensive pre-processing and hyper-parameter optimization. Extensive experiments on six Java projects demonstrate that FR-CNN consistently outperforms the developed traditional machine-learning baselines and achieves competitive performance against recent state-of-the-art approaches. While some methods, such as CBIL, achieve higher F1-scores on several individual projects, FR-CNN provides a more favorable balance between predictive performance, computational efficiency and model complexity. The additional efficiency analysis further shows that the proposed framework requires fewer trainable parameters, shorter training time and stable optimization behavior without overfitting, making it well suited for practical software fault-prediction applications.

The proposed FR-CNN is also consistent with the recent evolution of software fault-prediction research toward hybrid deep-learning architectures. Recent studies have shown that integrating complementary feature representations, dynamic classification strategies, attention mechanisms and heterogeneous deep learning models can substantially improve software fault prediction performance. In contrast to these recent approaches, the proposed FR-CNN employs a hierarchical stacking architecture that combines a fine-tuned FC-DNN and Random Forest using a CNN-based meta-learner, enabling the model to exploit complementary strengths of machine learning and deep learning for large-scale method-level software-fault prediction on the BugHunter dataset [21]-[24].

For future work, specific combinations of boosting or bagging algorithms with different structures of deep neural networks can be evaluated to potentially enhance the performance of the hybrid model. Also, it is valuable to conduct empirical studies on the rest of the projects at the method-level of the BugHunter dataset or other granularity levels, such as file-level and class-level for SFP. Moreover, we can consider the structure of CNN, LSTM, Dense and Dropout layers to detect software faults.

Data Availability

The datasets used in this study are publicly available through BugHunter (BugHunter Dataset-Mendeley Data).

Code Availability

To enhance the transparency, reproducibility and practical applicability of the proposed framework, the complete implementation of the FR-CNN approach has been made publicly available through an online GitHub repository (FR-CNN Hybrid approach). The repository includes the complete experimental workflow, encompassing data pre-processing, implementation of the proposed FR-CNN framework and implementations of the baseline deep-learning and traditional machine-learning models evaluated in this study. In addition, the publicly available BugHunter datasets used in the experiments, the optimal hyper-parameter configurations identified for each of the six projects and the corresponding pre-trained FC-DNN models are provided. To facilitate practical reuse, each pre-trained FC-DNN model is available in three complementary formats: the complete pre-trained model, the model architecture (JSON configuration) and the corresponding model weights. The implementation supports both direct loading of the complete pre-trained model and reconstruction of the model from its architecture and weights using TensorFlow/Keras. Since the FC-DNN base learners have already been trained using the optimal hyper-parameter configurations reported in Table 4, users can directly load the pre-trained models without repeating the computationally intensive training process. Consequently, reproducing the proposed FR-CNN framework requires zero additional training time for the FC-DNN stage, substantially reducing the computational cost while enabling researchers and practitioners to efficiently reproduce the reported experiments under the same experimental settings.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Funding Statement

The authors declare that this paper was not supported by any funding or financial assistance.

REFERENCES

[1] A. Arshad, S. Riaz, L. Jiao and A. Murthy, "Semi-supervised Deep Fuzzy C-mean Clustering for Software Fault Prediction," IEEE Access, vol. 6, pp. 25675–25685, 2018.

[2] F. Li, P. Yang, J. W. Keung, W. Hu, H. Luo and X. Yu, "Revisiting 'Revisiting Supervised Methods for Effort-aware Cross-project Defect Prediction'," IET Software, vol. 17, no. 4, pp. 472–495, 2023.

[3] E. N. Akimova et al., "A Survey on Software Defect Prediction Using Deep Learning," Mathematics, vol. 9, no. 11, p. 1180, 2021.

[4] S. K. Pandey et al., "Software Defect Prediction Using K-PCA and Various Kernel-based Extreme Learning Machine: An Empirical Study," IET Software, vol. 14, no. 7, pp. 768–782, 2020.

[5] C. Jin, "Cross-project Software Defect Prediction Based on Domain Adaptation Learning and Optimization," Expert Systems with Applications, vol. 171, p. 114637, 2021.

[6] J. Pachouly, S. Ahirrao, K. Kotecha, G. Selvachandran and A. Abraham, "A Systematic Literature Review on Software Defect Prediction Using Artificial Intelligence: Datasets, Data Validation Methods, Approaches and Tools," Engineering Applications of Artificial Intelligence, vol. 111, p. 104773, 2022.

[7] Y. Hassouneh et al., "Boosted Whale Optimization Algorithm with Natural Selection Operators for Software Fault Prediction," IEEE Access, vol. 9, pp. 14239–14258, 2021.

[8] S. Kumar and S. S. Rathore, Software Fault Prediction: A Road Map, eBook ISBN978-981-10-8715-8, Springer, 2018.

[9] Z. Li, X.-Y. Jing and X. Zhu, "Progress on Approaches to Software Defect Prediction," Iet Software, vol. 12, no. 3, pp. 161–175, 2018.

[10] X. Chen, Y. Zhao, Q. Wang and Z. Yuan, "MULTI: Multi-objective Effort-aware Just-in-time Software Defect Prediction," Information and Software Technology, vol. 93, pp. 1–13, 2018.

[11] K. Wang, L. Liu, C. Yuan and Z. Wang, "Software Defect Prediction Model Based on LASSO–SVM," Neural Computing and Applications, vol. 33, no. 14, pp. 8249–8259, 2021.

[12] S. K. Pandey, R. B. Mishra and A. K. Tripathi, "Machine Learning Based Methods for Software Fault Prediction: A Survey," Expert Systems with Applications, vol. 172, p. 114595, 2021.

[13] F. Khan et al., "Hyper-parameter Optimization of Classifiers, Using an Artificial Immune Network and Its Application to Software Bug Prediction," IEEE Access, vol. 8, pp. 20954–20964, 2020.

[14] C. Arun and C. Lakshmi, "Genetic Algorithm-based Oversampling Approach to Prune the Class Imbalance Issue in Software Defect Prediction," Soft Computing, vol. 26, no. 23, pp. 12915–12931, 2022.

[15] F. Matloob et al., "Software Defect Prediction Using Ensemble Learning: A Systematic Literature Review," IEEE Access, vol. 9, pp. 98754-98771, 2021.

[16] Ö. F. Arar and K. Ayan, "A Feature Dependent I Bayes Approach and Its Application to the Software Defect Prediction Problem," Applied Soft Computing, vol. 59, pp. 197–209, 2017.

[17] M. Singh and D. S. Salaria, "Software Defect Prediction Tool Based on Neural Network," Int. J. of Computer Applications, vol. 70, no. 22, 2013.

[18] Z. M. Zain et al., "Software Defect Prediction Harnessing on Multi 1-dimensional Convolutional Neural Network Structure," Computers, Materials and Continua, vol. 71, no. 1, p. 1521, 2022.

[19] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, [Online], Available: http://www.deeplearningbook.org, MIT Press, 2016.

[20] M. M. Ahsan, T. E Alam, T. Trafalis and P. Huebner, "Deep MLP-CNN Model Using Mixed-data to Distinguish between COVID-19 and Non-COVID-19 Patients," Symmetry, vol. 12, no. 9, p. 1526, 2020.

[21] A. Abdu et al., "Semantic and Traditional Feature Fusion for Software Defect Prediction Using Hybrid Deep Learning Model," Scientific Reports, vol. 14, no. 1, p. 14771, 2024.

[22] S. Kaliraj, V. G. P. Sahasranth and V. Sivakumar, "A Holistic Approach to Software Fault Prediction with Dynamic Classification," Automated Software Engineering, vol. 31, no. 2, p. 70, 2024.

[23] Q. Fei et al., "A Software Defect Prediction Method Using a Multivariate Heterogeneous Hybrid Deep Learning Algorithm," Computers, Materials, & Continua, vol. 82, no. 2, p. 3251, 2025.

[24] R. Malhotra and P. Singh, "DHG-BiGRU: Dual-attention based hierarchical gated BiGRU for software defect prediction," Information and Software Technology, vol. 179, p. 107646, 2025.

[25] J. S. Shirabad and T. J. Menzies, "The PROMISE Repository of Software Engineering Databases," School of Information Technology and Engineering, University of Ottawa, Canada, vol. 24, p. 3, 2005.

[26] M. Shepperd, Q. Song, Z. Sun and C. Mair, "Data Quality: Some Comments on the NASA Software Defect Datasets," IEEE Transactions on Software Engineering, vol. 39, no. 9, pp. 1208–1215, 2013.

[27] E. N. Akimova et al., "PyTraceBugs: A Large Python Code Dataset for Supervised Machine Learning in Software Defect Prediction," Proc. of the 2021 28th Asia-Pacific Software Engineering Conf. (APSEC), pp. 141–151, Taipei, Taiwan, 2021.

[28] R. Ferenc, P. Gyimesi, G. Gyimesi, Z. Tóth and T. Gyimóthy, "An Automatically Created Novel Bug Dataset and Its Validation in Bug Prediction," J. of Systems and Software, vol. 169, p. 110691, 2020.

[29] S. T. Cynthia et al., "Feature Transformation for Improved Software Bug Detection Models," Proc. of the 15th Innovations in Software Engineering Conf. (ISEC '22), Article no. 16, pp. 1–10, 2022.

[30] J. Jász, "The Effectiveness of Hidden Dependence Metrics in Bug Prediction," IEEE Access, vol. 12, pp. 77214–77225, 2024.
[31] S. S. Rathore and S. Kumar, "Linear and Non-linear Heterogeneous Ensemble Methods to Predict the Number of Faults in Software Systems," Knowledge-based Systems, vol. 119, pp. 232–256, 2017.
[32] S. S. Rathore and S. Kumar, "Towards an Ensemble Based System for Predicting the Number of Software Faults," Expert Systems with Applications, vol. 82, pp. 357–382, 2017.
[33] S. Wang and X. Yao, "Using Class Imbalance Learning for Software Defect Prediction," IEEE Transactions on Reliability, vol. 62, no. 2, pp. 434–443, 2013.
[34] Z. Sun, Q. Song and X. Zhu, "Using Coding-based Ensemble Learning to Improve Software Defect Prediction," IEEE Trans. on Systems, Man and Cybernetics, Part C, vol. 42, no. 6, pp. 1806–1817, 2012.
[35] S. S. Rathore and S. Kumar, "An Empirical Study of Ensemble Techniques for Software Fault Prediction," Applied Intelligence, vol. 51, no. 6, pp. 3615–3644, 2021.
[36] S. Moustafa et al., "Software Bug Prediction Using Weighted Majority Voting Techniques," Alexandria Engineering Journal, vol. 57, no. 4, pp. 2763–2774, 2018.
[37] I. H. Laradji, M. Alshayeb and L. Ghouti, "Software Defect Prediction Using Ensemble Learning on Selected Features," Information and Software Technology, vol. 58, pp. 388–402, 2015.
[38] T. Wang, Z. Zhang, X. Jing and L. Zhang, "Multiple Kernel Ensemble Learning for Software Defect Prediction," Automated Software Engineering, vol. 23, no. 4, pp. 569–590, 2016.
[39] S. K. Pandey et al., "BPDET: An Effective Software Bug Prediction Model Using Deep Representation and Ensemble Learning Techniques," Expert Systems with Applications, vol. 144, p. 113085, 2020.
[40] T.-Y. Yu, C.-Y. Huang and N. C. Fang, "Use of Deep Learning Model with Attention Mechanism for Software Fault Prediction," Proc. of the 2021 8th IEEE Int. Conf. on Dependable Systems and Their Applications (DSA), pp. 161–171, Yinchuan, China, 2021.
[41] H. Liang, Y. Yu, L. Jiang and Z. Xie, "Seml: A Semantic LSTM Model for Software Defect Prediction," IEEE Access, vol. 7, pp. 83812–83824, 2019.
[42] M. Alenezi, M. Zagane and Y. Javed, "Efficient Deep Features Learning for Vulnerability Detection Using Character N-gram Embedding," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 7, no. 01, pp. 25-38, 2021.
[43] U. Samal, "Hybrid Approach to Software Fault Prediction Using Particle Swarm Optimization and Fuzzy Time Series," Quality and Reliability Engineering International, vol. 41, no. 5, pp. 2005–2018, 2025.
[44] H. Turabieh, M. Mafarja and X. Li, "Iterated Feature Selection Algorithms with Layered Recurrent Neural Network for Software Fault Prediction," Expert Systems with Appl., vol. 122, pp. 27–42, 2019.
[45] H. Ji et al., "A New Weighted I Bayes Method Based on Information Diffusion for Software Defect Prediction," Software Quality Journal, vol. 27, no. 3, pp. 923–968, 2019.
[46] K. Zhu et al., "Software Defect Prediction Based on Enhanced Metaheuristic Feature Selection Optimization and a Hybrid Deep Neural Network," J. of Sys. and Software, vol. 180, p. 111026, 2021.
[47] B. S. Raghuwanshi and S. Shukla, "SMOTE Based Class-specific Extreme Learning Machine for Imbalanced Learning," Knowledge-Based Systems, vol. 187, p. 104814, 2020.
[48] N. Nnamoko and I. Korkontzelos, "Efficient Treatment of Outliers and Class Imbalance for Diabetes Prediction," Artificial Intelligence in Medicine, vol. 104, p. 101815, 2020.
[49] N. Limsettho et al., "Cross Project Defect Prediction Using Class Distribution Estimation and Oversampling," Information and Software Technology, vol. 100, pp. 87–102, 2018.
[50] O. Al Qasem, M. Akour and M. Alenezi, "The Influence of Deep Learning Algorithms Factors in Software Fault Prediction," IEEE Access, vol. 8, pp. 63945–63960, 2020.
[51] M. Adil, R. Ullah, S. Noor and N. Gohar, "Effect of Number of Neurons and Layers in an Artificial Neural Network for Generalized Concrete Mix Design," Neural Computing and Appl., pp. 1–9, 2022.
[52] J. Zhang et al., "Random-forests-based Network Intrusion Detection Systems," IEEE Trans. on Systems, Man and Cybernetics, Part C (Applications and Reviews), vol. 38, no. 5, pp. 649–659, 2008.
[53] L. Qiao, X. Li, Q. Umer and P. Guo, "Deep Learning Based Software Defect Prediction," Neurocomputing, vol. 385, pp. 100–110, 2020.
[54] A. B. Farid, E. M. Fathy, A. S. Eldin and L. A. Abd-Elmegid, "Software Defect Prediction Using Hybrid Model (CBIL) of Convolutional Neural Network (CNN) and Bidirectional Long Short-term Memory (Bi-LSTM)," PeerJ Computer Science, vol. 7, p. e739, 2021.

"FR-CNN: Software Fault Prediction Using a Novel Hybrid Hierarchical Model", M. M. Jouybari, A. Tajary, M. Fateh and E. Tahanian.

ملخص البحث:

يرتبط تطوير منتجاتٍ برمجيةٍ عالية الجودة ارتباطاً مباشراً بخلوّها من الأخطاء غير المكتشفة. وقد استُخدمت أساليب التّنبُّؤ بأخطاء البرمجيات لتقليل التّكاليف الإجمالية للاختبار والكشف عن الأخطاء المحتملة. واستُخدمت خوارزميات تعلّم آلي متنوعةٌ لهذا الغرض. إلّا أنّ أداءَها متفاوتٌ فيما يتعلّق بأخطاء البرمجيات.

يقترح هذا البحث نموذجاً هجيناً مبتكراً يعتمد على دمج ثلاث تقنياتٍ: شبكة عصبية عميقة متّصلة بالكامل (تم ضبط معاملاتها بدقّة)، وخوارزمية الغابة العشوائية، وشبكة عصبية تلافيفية. ومن خلال الاستفادة من نقاط القوّة النّاتجة عن دمج هذه التّقنيات، يهدف هذا النّموذج إلى تحسين أداء التّنبّؤ بأخطاء البرمجيات مع الحدّ من مشكلة الإفراط في التّخصيص، ممّا ينتج عنه إطار عملٍ قويٌ وفعّال.

وقد أجرينا دراساتٍ تجريبية موسّعة لتوضيح فعالية النّظام المقترح في التّنبّؤ بالأخطاء عبر ستة مشاريع من مجموعة بيانات (BugHunter). وأظهرت النّتائج التّجريبية أنّ النّموذج المقترح يحقّق أداءً تنافسياً في التّنبّؤ بالأخطاء، مع تطلّبه لتعقيد حسابي أقلّ وزمن تدريب أقصر. وقد تفوّق على أساليب التّعلُّم الآلي التّقليدية في عددٍ من مؤشّرات الأداء على مجموعة البيانات المذكورة.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

NETWORK PERFORMANCE OF TWO-WAY RELAYING WITH FOUNTAIN CODING IN WIRELESS SENSOR NETWORKS UNDER CO-CHANNEL INTERFERENCE

Duy-Hung Ha[1], Kien-Duc Tran[2], Dac-Binh Ha[3], Hien-Luong Nguyen[4], The-Anh Ngo[5] and Quoc-Hao Truong[6]

(Received: 22-May-2026, Revised: 4-Aug.-2026, Accepted: 20-Aug.-2026)

ABSTRACT

This paper investigates the performance of a two-way relay network (TWRN) for wireless sensor networks (WSNs) that combines fountain coding (FC) and digital network coding (DNC) in the presence of co- channel interference (CCI). A three-phase FC–DNC transmission protocol (SM-3P-FC) is considered and compared with the conventional four-phase fountain-coded protocol (SM-4P-FC). To isolate the contribution of fountain coding, an additional same-budget three-phase uncoded fixed-repetition benchmark (SM-3P-NoFC) is introduced, in which the Q transmission opportunities are distributed among H distinct uncoded packets. Under independent block-Rayleigh fading, closed-form analytical expressions for the outage probabilities at the two terminal nodes are derived by accounting for CCI at both the relay and the source nodes. Based on the derived outage probabilities, a bandwidth-normalized sum-throughput expression is further developed to characterize the trade-off between the target spectral efficiency and the probability of successful message recovery. Monte Carlo simulations based on 10^6 independently generated channel realizations per plotted point show close agreement with the derived analytical expressions over the evaluated parameter settings. The numerical results show that the proposed SM-3P-FC scheme achieves lower outage probabilities and higher bidirectional sum throughput than both the SM-3P-NoFC and SM-4P-FC benchmarks under the considered system settings. The additional CCI-power comparison further shows that the outage probabilities continue to decrease when interference is absent or when its power remains fixed, whereas proportional desired/interference power scaling produces non-zero outage floors.

KEYWORDS

Fountain coding, Digital network coding, Two-way relay network, Three-phase transmission, Co-channel interference, Outage probability, Sum throughput, Rayleigh fading.

1. INTRODUCTION

In dense or infrastructure-limited wireless sensor networks (WSNs), two sensor-cluster heads or aggregation nodes may need to exchange control or sensed-data packets through an intermediate relay, because the direct link is unavailable or insufficiently reliable. Such nodes typically operate under limited energy, transmit power, memory and processing capability. Therefore, reducing the number of transmission phases and limiting packet-specific retransmissions are important for transmission latency and energy efficiency in WSN deployments [24][39]. In the link-level abstraction adopted in this work, S_1 and S_2 may represent two sensor-cluster heads or aggregation nodes, whereas R represents an intermediate relay or gateway.

A two-way relay network (TWRN) is attractive in this setting, because it enables two end nodes to exchange information through a common relay while using fewer radio resources than two independent one-way relay transmissions [1]–[4]. However, frequency reuse among neighboring sensor clusters or communication devices may generate co-channel interference (CCI) at both the relay and the terminal nodes. This interference reduces the received signal-to-interference- plus-noise ratio

1. D.-H. Ha (Corresponding Author) is with Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam. Email: haduyhung@tdtu.edu.vn
2. K.-D. Tran is with Faculty of Electrical and Electronics Engineering, Ton Duc Thang University, Ho Chi Minh City, Vietnam. Email: trankienduc.st@tdtu.edu.vn
3. D.-B. Ha is with School of Engineering and Technology, Duy Tan University, Da Nang, Vietnam. Email: hadacbinh@duytan.edu.vn
4. H.-L. Nguyen is with Department of Human Resources and Administration, Ton Duc Thang University, Ho Chi Minh City, Vietnam. Email: nguyenhienluong@tdtu.edu.vn
5. T.-A. Ngo is with Faculty of Engineering and Technology, Campus in Ho Chi Minh City, University of Transport and Communications, Hanoi, Vietnam. Email: anhnt_ph@utc.edu.vn
6. Q.-H. Truong is with Go Dau Infrastructure Center – VNPT. Email: truongquochao9792@gmail.com

and may cause packet-decoding failures, particularly in dense or power-limited WSN deployments [6]–[8].

Digital network coding (DNC) and fountain coding (FC) address different, but complementary, aspects of this problem. DNC allows the relay to combine two successfully decoded packets through a packet-level XOR operation and thereby complete the bidirectional exchange in three transmission phases instead of four. FC improves packet-level recovery by allowing a destination to reconstruct a source block after collecting a sufficient number of encoded packets, rather than requiring every particular encoded packet to be successfully delivered [3][20]. Consequently, the joint use of TWR, DNC and FC is motivated by the need to reduce the exchange duration while maintaining reliable packet recovery in the presence of CCI.

Accordingly, this work focuses on the packet-level outage probability and bandwidth-normalized bidirectional sum throughput of a terrestrial FC–DNC-assisted two-way relay link operating under CCI. The WSN context is represented through a physical- and link-layer abstraction of resource-constrained bidirectional packet exchange, while higher-layer routing, medium-access- control and network-management functions are beyond the scope of this study. This abstraction isolates the trade-off among transmission-phase reduction, fountain-code recovery and co-channel interference in a WSN-motivated two-way relay segment.

Existing studies related to the present work can be broadly classified into three research directions. The first research direction concerns interference-aware and multi-user two-way relaying. Nguyen and Do [6] analyzed a two-way NOMA network in which CCI affects both the relay and the destination nodes, while considering full-duplex and half-duplex operation together with successive interference cancellation. Nam et al. [7] investigated an interference-cancellation-aided cognitive two-way relay network with primary transmit-antenna selection/selection combining and secondary partial relay selection. Ozduran and Soleimani-Nasab [8] further studied multi-user two-way relaying in the presence of CCI and evaluated its impact on the achievable sum-rate performance.

For multi-user amplify-and-forward (AF) bidirectional relaying, opportunistic source-pair selection (OSPS) was introduced to mitigate inter-cell interference (ICI) by activating an appropriate source pair according to a sum-rate-based criterion [31]. This OSPS approach was subsequently analyzed in greater detail for multi-user two-way relay networks, where maximum-sum-capacity source-pair selection was shown to provide multi-user diversity and improved performance relative to simultaneous transmission and max–min selection [32]. Its robustness under imperfect channel state information (CSI) resulting from feedback delay and channel-estimation error (CEE) was further investigated in [33].

Related studies have also examined more advanced interference-aware two-way relay architectures. Ozduran et al. [34] analyzed power-domain NOMA for full-duplex two-way relaying and evaluated outage probability, error probability, ergodic rate and throughput while optimizing the transmit powers, power-allocation coefficients and relay position. Ozduran [35] investigated the secrecy achievable rate of an illegitimate full-duplex DF NOMA two-way relay network affected by CCI and friendly jamming. In addition, the impact of CCI on sum-rate-based opportunistic relay selection in dual-hop multiple full-duplex bidirectional relaying was studied in [36].

These studies demonstrate that source-pair scheduling, relay selection, NOMA, full-duplex operation and interference-management mechanisms can significantly improve the performance of bidirectional relay networks. However, their primary objectives are multi-user interference coordination, full-duplex NOMA, relay selection or physical-layer security. They do not jointly consider packet-level fountain coding and digital network coding, nor do they provide a matched SM-3P-FC/SM-4P-FC analysis with explicit CCI at the relay and both terminal nodes.

The second research direction employs fountain coding in cooperative and two-way relay networks. Fountain-coded cooperative schemes allow a destination to recover a source message after collecting a sufficient number of encoded packets, thereby improving reliability under packet loss and fading. Recent extensions have integrated fountain coding with cognitive-radio operation, reconfigurable intelligent surfaces and radio-frequency energy harvesting [20]. Nevertheless, these systems address different network architectures and do not provide a matched comparison between three-phase and four-phase transmission under explicit co-channel interference at all receiving nodes.

The third research direction focuses on fountain-coded and network-coded two-way relaying. Hau et al. investigated a secure multi-cast two-way relaying system employing fountain coding, three-phase digital network coding and randomize-and-forward transmission [26]. Their analysis considered outage probability, intercept probability, system-outage probability and system-intercept probability over Rayleigh-fading channels.

Toan et al. considered a secure two-way hybrid satellite–terrestrial relay system employing fountain coding, successive interference cancellation, digital network coding, partial relay selection and cooperative jamming [27]. In that scheme, the satellite and ground users simultaneously transmit encoded packets to the relay stations in the first phase, after which a selected relay performs an XOR operation and broadcasts the resulting packet in the second phase.

Recent studies published in 2024–2026 have further extended fountain coding to a variety of relay architectures. Nguyen et al. [20] investigated a fountain-coded two-way cognitive relaying network employing a reconfigurable intelligent surface and radio-frequency energy harvesting. Hau et al. [26] considered a secure multi-cast two-way relaying system combining fountain coding, three-phase digital network coding and randomize-and-forward transmission. Other recent studies employed fountain coding in hybrid satellite–terrestrial NOMA relay systems with RIS assistance or partial relay selection and physical-layer security [15].

More recently, Toan et al. [44] investigated fountain-coded hybrid satellite–terrestrial relay multi-cast under co-channel interference. They derived exact outage and system-outage probability expressions and formulated a joint time- and power-allocation problem to improve system reliability. Although that work explicitly considers both fountain coding and CCI, it focuses on one-way satellite-assisted multicast rather than terrestrial bidirectional relaying with DNC.

Among these works, the scheme in [26] is the closest to the present study, because it already combines two-way relaying, fountain coding and three-phase digital network coding. This confirms that these individual techniques have already been combined in related two-way relay architectures. Nevertheless, the existing studies focus on secure multicast, cognitive-radio, RIS-assisted, energy-harvesting or hybrid satellite–terrestrial systems. They do not provide a unified closed- form analysis of a terrestrial two-way relay network in which co-channel interference explicitly affects the relay and both terminal nodes. Moreover, they do not provide matched comparisons among SM-3P-FC, SM-4P-FC and a same-budget SM-3P-NoFC benchmark under identical channel, interference and power conditions.

From a WSN perspective, the unresolved issue is the absence of a unified evaluation of the three requirements that motivate the considered link-level scenario: reduced-phase bidirectional packet exchange, flexible source-block recovery and explicit CCI at every receiving node. Interference-aware two-way relaying studies generally omit packet-level fountain coding, whereas recent fountain-coded relay studies mainly focus on cognitive-radio, RIS-assisted, energy-harvesting, security or hybrid satellite–terrestrial architectures. Consequently, their results do not directly quantify the separate reliability and throughput contributions of fountain coding and three- phase transmission under matched terrestrial channel, interference, power and packet-budget conditions.

Accordingly, the novelty of the present work does not lie in introducing fountain coding, digital network coding or three-phase relaying individually. Rather, it lies in the unified closed- form performance analysis of a terrestrial SM-3P fountain-coded DNC two-way relay network with explicit co-channel interference at the relay and both terminal nodes. The developed framework derives the end-node outage probabilities and a bandwidth-normalized bidirectional sum-throughput expression and provides matched comparisons among the proposed SM-3P-FC scheme, the conventional SM-4P-FC scheme and a same-budget SM-3P-NoFC benchmark under identical channel, interference and power conditions, with the same source-block size and transmission budget. These comparisons separately identify the reliability and throughput contributions of fountain coding and three-phase transmission.

Objective and Contributions

Motivated by the above WSN link-level scenario, this paper develops and analyzes a terrestrial three-phase TWR scheme that jointly employs fountain coding and digital network coding under co-channel interference. The primary objective is to quantify the separate and combined effects of

fountain-code-based recovery and three-phase transmission on the outage probability and bandwidth-normalized bidirectional sum throughput when the relay and both terminal nodes are exposed to CCI. The proposed SM-3P-FC scheme is analytically and numerically compared with both the conventional SM-4P-FC protocol and a same-budget three-phase uncoded fixed-repetition benchmark, denoted by SM-3P-NoFC. The main contributions of this work are summarized as follows:

- We develop an SM-3P-FC architecture that jointly incorporates fountain coding, packet-level digital network coding and co-channel interference in a two-way relay network. The considered model explicitly accounts for interference at the relay and both terminal nodes.
- We derive closed-form expressions for the per-round packet-delivery probabilities and the resulting outage probabilities at both terminal nodes over independent block-Rayleigh fading channels. The analysis explicitly captures the transmit powers, path-loss parameters, number of interferers, relay position, fountain-code recovery threshold H and maximum number of exchange rounds Q.
- Based on the derived outage probabilities, we develop a bandwidth-normalized bidirectional sum-throughput expression. This metric characterizes the trade-off among the target spectral efficiency, message-recovery probability and fountain-code parameters and reveals the existence of an optimum target-spectral efficiency for each transmission protocol.
- We conduct extensive Monte Carlo simulations to validate the analytical outage-probability and bidirectional sum-throughput expressions. The proposed SM-3P-FC scheme is evaluated against both the conventional SM-4P-FC protocol and a newly introduced same- budget SM-3P-NoFC benchmark. These comparisons separately reveal the performance contributions of fountain coding and the reduction in the number of transmission phases. The numerical results also quantify the effects of normalized transmit power, interference- power scaling, relay placement, H, Q and the number of co-channel interferers and provide engineering guidelines for interference-limited WSN and IoT deployments.

Table 1 summarizes the main differences between the present study and representative interference-aware, fountain-coded and network-coded two-way relaying works. The comparison shows that the individual components of the considered three-phase architecture have appeared in related contexts, whereas the present contribution focuses on their unified terrestrial CCI analysis and matched evaluations among the SM-3P-FC, SM-4P-FC and same-budget SM-3P-NoFC schemes.

Table 1. Comparison with representative related works.

Ref.	System and interference context	FC	Coding and transmission protocol	Metrics	Main distinction
Ref. [6]	Two-way NOMA with CCI at the relay and destination nodes	×	FD/HD relaying with perfect or imperfect SIC	OP	Analyzes the joint effects of CCI, duplex mode and SIC, without fountain coding or a matched SM-3P-FC/SM-4P-FC comparison.
Ref. [7]	Underlay cognitive two-way relay network with interference cancellation	×	DNC, primary TAS/SC and secondary partial relay selection	OP	Studies interference management and relay selection in cognitive two-way relaying, but does not employ fountain coding.
Ref. [20]	Cognitive two-way relaying with RIS and RF energy harvesting	✓	Three-phase DNC with conventional and early-stopping FC schemes	OP, SOP and average transmissions	Focuses on cognitive-radio interference constraints, RIS and energy harvesting rather than explicit CCI at every receiving node.
Ref. [26]	Secure multi-cast two-way relaying in the presence of a passive eavesdropper	✓	Three-phase FC–DNC with randomize-and-forward transmission	OP, IP, SOP, and SIP	Combines FC and three-phase DNC for secure multi-cast and physical-layer security, rather than terrestrial CCI explicitly affecting the relay and both terminal nodes.
Ref. [27]	Secure hybrid satellite–errestrial two-way relaying with an eavesdropper and cooperative jammer	✓	Two-phase FC–DNC with SIC, partial relay selection and cooperative jamming	OP, SOP, IP, and SIP	Analyzes the security–reliability trade-off in a hybrid satellite–terrestrial system, rather than conventional terrestrial CCI at the relay and both terminal nodes.
This paper	Terrestrial TWRN with explicit CCI at the relay and both terminals	✓	Three-phase FC–DNC; matched SM-4P-FC and same-budget SM-3P-NoFC benchmarks	OP and bidirectional sum throughput	Derives closed-form end-node OP and bandwidth-normalized bidirectional sum throughput; provides matched comparisons among SM-3P-FC, SM-4P-FC and a same-budget SM-3P-NoFC benchmark under identical channel, interference and power conditions.

Abbreviations

Abbreviation	Definition	Abbreviation	Definition
TWR	Two-way relaying	RF	Radio frequency
TWRN	Two-way relay network	IP	Intercept probability
WSN	Wireless sensor network	SOP	System-outage probability
IoT	Internet of Things	SIP	System-intercept probability
SM-3P	Three-phase transmission protocol	TAS/SC	Transmit-antenna selection/selection combining
SM-4P	Four-phase transmission protocol	AF	Amplify-and-forward
FC	Fountain coding	DF	Decode-and-forward
DNC	Digital network coding	ICI	Intercell interference
CCI	Co-channel interference	OSPS	Opportunistic source-pair selection
OP	Outage probability	CSI	Channel state information
SINR	Signal-to-interference-plus-noise ratio	CEE	Channel-estimation error
CDF	Cumulative distribution function	MAC	Multiple-access channel
PDF	Probability density function	BC	Broadcast channel
MC	Monte Carlo	LDPC	Low-density parity-check
AWGN	Additive white Gaussian noise	OFDM	Orthogonal frequency-division multiplexing
NOMA	Non-orthogonal multiple access	NoFC	Without fountain coding
FD/HD	Full-duplex/half-duplex	SM-3P-NoFC	Three-phase DNC benchmark without fountain coding
SIC	Successive interference cancellation	SM-3P-FC	Three-phase FC–DNC transmission protocol
RIS	Reconfigurable intelligent surface	SM-4P-FC	Four-phase fountain-coded transmission protocol

Main Notations

Notation	Definition
S_1, S_2	Terminal/source nodes
R	Intermediate relay node
I_k	The k-th co-channel interferer
K	Number of co-channel interferers
m_1, m_2	Original source messages generated by S_1 and S_2, respectively
X_1, X_2	Packet-level payloads transmitted by S_1 and S_2, respectively, in a representative exchange round; fountain-coded for the FC schemes and uncoded for the NoFC benchmark
$X_3 = X_1 \oplus X_2$	Packet-level network-coded packet generated at R when both source packets are successfully decoded
$\mathcal{M}(\cdot)$	Modulation mapping from a packet-level payload to its transmitted symbol
$x_i = \mathcal{M}(X_i), i \in \{1,2,3\}$	Unit-energy modulated symbol associated with packet-level payload X_i
u_k	Unit-energy interference symbol transmitted by I_k
n_b	AWGN sample at receiving node $b \in \{R, S_1, S_2\}$
σ_0^2	AWGN power at each receiving node
P_S	Common transmit power of S_1, S_2 and R
P_I	Transmit power of each co-channel interferer
$\Psi_S = P_S/\sigma_0^2$	Normalized desired-node transmit power
$\Psi_I = P_I/\sigma_0^2$	Normalized co-channel-interference power
h_{ab}	Small-scale fading coefficient of link $a \to b$
$g_{ab} = \lvert h_{ab} \rvert^2$	Channel power gain of link $a \to b$
d_{ab}	Distance between nodes a and b
$\lambda_{ab} = d_{ab}^{\beta}$	Exponential rate parameter of the channel power gain on link $a \to b$
β	Path-loss exponent
C_{th}	Target spectral-efficiency threshold
$\rho_{\text{th}} = 2^{3C_{\text{th}}} - 1$	SINR threshold for the SM-3P protocol
$\xi_{\text{th}} = 2^{4C_{\text{th}}} - 1$	SINR threshold for the SM-4P protocol
ω_1, ω_2	Auxiliary parameters used in the SM-3P probability analysis
ϑ_1, ϑ_2	Auxiliary parameters used in the SM-4P probability analysis
H	Number of source packets and, under the ideal zero-overhead abstraction, the minimum recovery threshold
H_{eff}	Effective practical recovery threshold accounting for fountain-code overhead
Q	Maximum number of packet-exchange rounds
N_i	Number of fountain-coded packets successfully received by terminal S_i over Q exchange rounds
$\theta_{X_i}^{(p)}$	Per-round successful delivery probability of packet-level payload X_i under protocol p
$\text{OP}_{S_i}^{(p)}$	Outage probability at terminal S_i under protocol p
$\eta_S^{(p)}$	Bandwidth-normalized throughput at terminal S_i under protocol p
$\eta_{sum}^{(p)}$	Bidirectional bandwidth-normalized sum throughput under protocol p

p	Protocol index, $p \in$ {SM-3P, SM-4P}
N_{MC}	Number of independently generated channel realizations for each plotted point
$\mathcal{CN}(\mu, \sigma^2)$	Circularly symmetric complex Gaussian distribution with mean μ and variance σ^2
$\mathrm{Exp}(\lambda)$	Exponential distribution with rate parameter λ
$L = \lfloor Q/H \rfloor$	Minimum number of repetitions assigned to each uncoded source packet
$r = Q - LH$	Number of uncoded source packets assigned one additional repetition

The remainder of this paper is organized as follows. Section 2 presents the system and channel models and derives the outage-probability and sum-throughput expressions for the SM-3P-FC, SM-4P-FC and SM-3P-NoFC schemes. Section 3 provides the analytical and Monte Carlo simulation results and discusses the corresponding engineering design implications. Finally, Section 4 concludes the paper and outlines future research directions.

2. System Model and Performance Analysis

2.1 Common System Model and Three-Phase DNC Protocol (SM-3P)

The network topology, node locations, transmit-power definitions, CCI model, channel distributions and block-fading assumptions introduced in this sub-section are common to the SM-3P-FC, SM-4P-FC and SM-3P-NoFC schemes. The schemes differ only in their packet-generation and recovery rules, relay-processing operations and transmission-phase schedules. Accordingly, the common system assumptions are stated once in this sub-section, whereas Sub-sections 2.3 and 2.5 subsequently present only the protocol-specific operations and performance expressions.

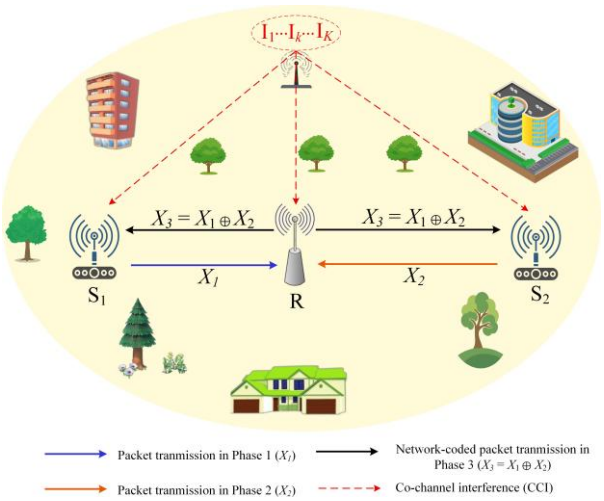


Figure 1. TWR model using DNC operating in a co-channel interference environment.

For clarity, Figure 1 illustrates the packet-level information flow of the three-phase protocol. The relay performs the XOR operation at the packet level as $X_3 = X_1 \oplus X_2$. Before transmission over the wireless channel, each packet-level payload X_i is mapped to the corresponding unit-energy modulated symbol $x_i = \mathrm{M}(X_i)$, $i \in \{1,2,3\}$, as used in the subsequent signal model.

The considered three-phase two-way relay network with DNC, hereafter denoted by SM-3P, consists of two source nodes, S_1 and S_2 and an intermediate relay R. Direct communication between S_1 and S_2 is unavailable (e.g., due to being out of range or obstructed), so both sources rely on R [25]. The network also includes K co-channel interferers affecting the network, denoted by $I_1, I_2, \ldots, I_K$. Such co-channel interference arises from frequency reuse, in which the spectrum is reused to improve spectral efficiency.

For the fountain-coded schemes, S_1 and S_2 generate fountain-coded packets and exchange them over successive packet-exchange rounds. Let X_1 and X_2 denote the packet-level payloads transmitted by S_1 and S_2, respectively, in a representative exchange round. Their corresponding unit-energy modulated symbols are denoted by $x_1 = M(X_1)$ and $x_2 = M(X_2)$. For notational simplicity, the exchange-round index is omitted from the instantaneous signal model.

Let m_1 and m_2 denote the original source messages generated by S_1 and S_2, respectively. To recover

m_2, terminal S_1 must successfully collect at least H fountain-coded packets generated by S_2. Similarly, S_2 must collect at least H fountain-coded packets generated by S_1 to recover m_1. The exchange proceeds for at most Q packet-exchange rounds, after which a terminal is considered to be in outage if it has collected fewer than H required encoded packets.

The present analysis adopts an ideal packet-level fountain-coding model. Specifically, H denotes the number of source packets and, under the adopted zero-overhead abstraction, also the minimum number of correctly received encoded packets assumed to be sufficient for successful source-message recovery, whereas Q denotes the maximum number of packet-exchange rounds.

In a practical fountain-code implementation, successful decoding may require more than H correctly received encoded packets because of finite-length and degree-distribution overhead. At the present packet-level abstraction, this overhead may be approximated by replacing H with an effective recovery threshold $H_{\text{eff}} > H$. The excess $H_{\text{eff}} - H$ depends on the selected code construction, degree distribution, source-block length and decoding algorithm. The degree distribution also affects the sparsity of the decoding graph and therefore the number of decoding operations, memory usage, processing latency and energy consumption at resource-constrained WSN nodes [28]-[29].

This effective-threshold substitution is only a packet-level approximation of finite-length fountain decoding. A code-specific analysis would additionally require a decoding-success function that depends on the number of correctly received packets, source-block length, degree distribution, code construction and decoding algorithm. Under the effective-threshold approximation, the predicted outage probability generally increases and the effective throughput decreases.

If R successfully decodes only packet X_1, but not packet X_2, it forwards X_1 to S_2 in the third time slot by transmitting the corresponding modulated symbol $x_1=M(X_1)$. Conversely, if R successfully decodes X_2, but not X_1, it forwards X_2 to S_1 by transmitting $x_2 =M(X_2)$.

Under the DNC method, the exchange over three time slots operates as follows. In the first two time slots, S_1 and S_2 transmit $x_1 = M(X_1)$ and $x_2 = M(X_2)$ to R, respectively; the transmission order is immaterial and does not affect performance. The relay then attempts to decode the packet-level payloads X_1 and X_2. If both packets are successfully decoded, R forms the packet-level network-coded packet, maps it to the unit-energy modulated symbol $x_3 = M(X_3)$ and broadcasts x_3 to S_1 and S_2 in the third time slot.

$$X_3 = X_1 \oplus X_2,$$

Protocol-specific differences: Within the unified system model, the SM-3P-FC and SM-3P-NoFC schemes employ the same three-phase physical transmission schedule and the same packet-level DNC operation. In the first phase, S_1 transmits $x_1 = M(X_1)$ to R and in the second phase, S_2 transmits $x_2 = M(X_2)$ to R. In the third phase, if both packets X_1 and X_2 are successfully decoded, the relay forms $X_3 = X_1 \oplus X_2$ and broadcasts $x_3 = M(X_3)$ to both terminal nodes. If only one packet is successfully decoded, the relay forwards that packet to its intended destination. If neither packet is decoded, the relay remains silent during the third phase.

Table 2. Transmission-phase schedules of the considered schemes.

Scheme	Phase 1	Phase 2	Relay phase(s)
SM-3P-FC	$S_1 \to R$	$S_2 \to R$	Phase 3: $R \to \{S_1, S_2\}$ using $X_3 = X_1 \oplus X_2$
SM-3P-NoFC	$S_1 \to R$	$S_2 \to R$	Phase 3: $R \to \{S_1, S_2\}$ using $X_3 = X_1 \oplus X_2$
SM-4P-FC	$S_1 \to R$	$S_2 \to R$	Phase 3: $R \to S_2$; Phase 4: $R \to S_1$

The difference between SM-3P-FC and SM-3P-NoFC lies in the packet-generation and source-message-recovery rules rather than in the physical transmission schedule. In SM-3P-FC, each source generates fountain-coded packets and the destination can recover the source message after collecting any H successfully received encoded packets over at most Q exchange rounds.

In SM-3P-NoFC, each source instead transmits H distinct uncoded packets according to a fixed repetition schedule and successful recovery requires every distinct source packet to be received at least once. By contrast, SM-4P-FC uses the same network topology, transmit-power definitions, CCI model and channel assumptions, but it forwards the two communication directions separately over four phases:

$S_1 \longrightarrow R$, $R \longrightarrow S_2$, $S_2 \longrightarrow R$ and $R \longrightarrow S_1$. Therefore, SM-4P-FC does not employ the packet-level XOR broadcast used by the two SM-3P schemes. The transmission-phase schedules of the three considered schemes are summarized in Table 2.

Practical implementation considerations: In an actual implementation, each destination must identify and buffer the successfully received encoded packets and then execute the selected fountain-decoding algorithm. The packet headers or coding metadata must also provide sufficient information for the decoder to determine the relationships among the received encoded packets. At the relay, the received signals from the two sources must be demodulated and decoded at the packet level. The successfully recovered encoded packets are then stored when necessary, combined through packet-level XOR when both packets are available, remodulated and broadcast to the terminal nodes.

These operations introduce packet-header overhead, signaling, buffering, memory usage, processing delay, decoding complexity and energy consumption. Such implementation costs can be significant for resource-constrained WSN devices. They are not included in the present outage analysis, the purpose of which is to isolate the effects of fading, co-channel interference, fountain-code recovery parameters and the number of transmission phases.

Transmission-phase and reciprocity assumptions: Unlike the two-phase MAC/BC information exchange considered in Ref. [8], the adopted SM-3P protocol employs two orthogonal source-to-relay time slots followed by one relay broadcast time slot. Specifically, S_1 transmits to R in the first time slot, S_2 transmits to R in the second time slot and R broadcasts or forwards the recovered packet information in the third time slot. Therefore, the first two time slots do not constitute a simultaneous two-user MAC phase.

The forward and reverse links between S_i and R, $i \in \{1,2\}$, have the same propagation distance and hence share the same large-scale channel parameter, i.e., $\lambda_{S_iR} = \lambda_{RS_i} = d_{S_iR}^{\beta}$.

For notational simplicity, λ_{S_iR} is used for both directions. However, the transmissions $S_i \longrightarrow R$ and $R \longrightarrow S_i$ occur in different time slots and their small-scale fading coefficients are modeled as independent realizations. Consequently, instantaneous channel reciprocity is not assumed, although the forward and reverse links have the same marginal distribution.

Signal-normalization and channel assumptions: The modulated symbols $x_i = \mathcal{M}(X_i)$, $i \in \{1,2,3\}$ and the interference symbols u_k are assumed to have unit average energy, i.e., $\mathbb{E}\left[\left|x_j\right|^2\right] = 1, j \in \{1,2,3\}$, and $\mathbb{E}[|u_k|^2] = 1, k \in \{1, \dots, K\}$.

The interference symbols are mutually independent and are independent of the desired transmitted symbols, channel coefficients and receiver noise. The AWGN sample at receiving node $b \in \{R, S_1, S_2\}$ is modeled as: $n_b \sim \mathcal{CN}(0, \sigma_0^2)$.

All desired and interfering links experience independent block-Rayleigh fading [37]. More specifically, $h_{S_iR} \sim \mathcal{CN}\left(0, \lambda_{S_iR}^{-1}\right), h_{RS_i} \sim \mathcal{CN}\left(0, \lambda_{S_iR}^{-1}\right), i \in \{1,2\}$, whereas $h_{I_kb} \sim \mathcal{CN}\left(0, \lambda_{I_kb}^{-1}\right), k \in \{1, \dots, K\}, b \in \{R, S_1, S_2\}$.

Accordingly, each channel power gain satisfies $g_{ab} = |h_{ab}|^2 \sim \text{Exp}(\lambda_{ab})$.

Each channel coefficient remains constant within one time slot and is independently regenerated across different time slots and packet-exchange rounds. The fading coefficients of distinct desired and interfering links are mutually independent. The node locations, transmit powers and link-specific marginal channel distributions remain unchanged over the Q exchange rounds.

2.1.1 First Time Slot

The signal received at R when S_1 transmits the unit-energy modulated symbol $x_1 = \mathcal{M}(X_1)$ associated with packet X_1 in the presence of the co-channel interferers is given by [7]:

$$y_{S_1 \to R} = \sqrt{P_S} h_{S_1R} x_1 + \sum_{k=1}^{K} \sqrt{P_I} h_{I_kR} u_k + n_R, \quad (1)$$

where P_S denotes the common transmit power used by S_1, S_2 and R, whereas P_I denotes the common

transmit power of each co-channel interferer. This equal-power assumption is applied within each transmitter group to simplify the analysis; no equality between P_S and P_I is imposed in the analytical derivations. The coefficients h_{S_1R} and h_{I_kR} denote the Rayleigh fading coefficients of the $S_1 \to R$ and $I_k \to R$ links, respectively, where $k \in \{1,2,\dots,K\}$.

From Eq. (1), the instantaneous signal-to-interference-plus-noise ratio (SINR) at R in the first time slot is expressed as:

$$\varphi_{S_1R} = \frac{P_S\left|h_{S_1R}\right|^2}{\sum_{k=1}^{K} P_I\left|h_{I_kR}\right|^2 + \sigma_0^2} = \frac{\Psi_S g_{S_1R}}{\sum_{k=1}^{K} \Psi_I g_{I_kR} + 1}, \tag{2}$$

where $g_{S_1R} = \left|h_{S_1R}\right|^2, g_{I_kR} = \left|h_{I_kR}\right|^2, \Psi_S = P_S/\sigma_0^2$ and $\Psi_I = P_I/\sigma_0^2$.

For subsequent averaging over the desired and interfering channel gains, the marginal CDFs and PDFs of g_{S_1R} and g_{I_kR} are given by:

$$\begin{aligned} F_{g_{S_1R}}(x) &= 1 - \exp\left(-\lambda_{S_1R}x\right), F_{g_{I_kR}}(x) = 1 - \exp\left(-\lambda_{I_kR}x\right), \\ f_{g_{S_1R}}(x) &= \lambda_{S_1R}\exp\left(-\lambda_{S_1R}x\right), f_{g_{I_kR}}(x) = \lambda_{I_kR}\exp\left(-\lambda_{I_kR}x\right). \end{aligned} \tag{3}$$

where $\lambda_{S_1R} = \left(d_{S_1R}\right)^{\beta}$ and $\lambda_{I_kR} = \left(d_{I_kR}\right)^{\beta}$. Here, d_{S_1R} and d_{I_kR} denote the corresponding link distances and β is the path-loss exponent, which typically satisfies $2 \le \beta \le 6$, depending on the propagation environment [38]. It should be emphasized that Eq. (3) characterizes the marginal distributions of the individual desired and interfering channel power gains; it is not the CDF of the SINR in Eq. (2). The aggregate CCI term in Eq. (2), $\sum_{k=1}^{K} \Psi_I g_{I_kR}$, is explicitly retained in the SINR expression and is subsequently accounted for by averaging over the joint distribution of the K independent interference gains, as shown in Eqs. (21-23). Accordingly, the instantaneous channel capacity at R (in this slot) is:

$$C_{S_1R} = \frac{1}{3}\log_2\left(1 + \varphi_{S_1R}\right) = \frac{1}{3}\log_2\left(1 + \frac{\Psi_S g_{S_1R}}{\sum_{k=1}^{K} \Psi_I g_{I_kR} + 1}\right). \tag{4}$$

Note: For a fair comparison between the three-phase TWR model (DNC) and the conventional four-phase TWR model, assume that the total transmission time per packet exchange is the same (normalized to one time unit) in both models. Since the SM-3P uses three time slots, each slot occupies one third of that time; this explains the factor $\frac{1}{3}$ preceding the logarithm in Eq. (4).

2.1.2 Second Time Slot

Similarly, when S_2 transmits the unit-energy modulated symbol $x_2 = \mathcal{M}(X_2)$ associated with packet X_2, the signal received at R is given by [7]:

$$y_{S_2 \to R} = \sqrt{P_S}h_{S_2R}x_2 + \sum_{k=1}^{K} \sqrt{P_I}h_{I_kR}u_k + n_R, \tag{5}$$

where h_{S_2R} denotes the channel coefficient of the $S_2 \to R$ link.

Hence, the instantaneous SINR at R in the second time slot is expressed as:

$$\varphi_{S_2R} = \frac{P_S\left|h_{S_2R}\right|^2}{\sum_{k=1}^{K} P_I\left|h_{I_kR}\right|^2 + \sigma_0^2} = \frac{\Psi_S g_{S_2R}}{\sum_{k=1}^{K} \Psi_I g_{I_kR} + 1}, \tag{6}$$

where $g_{S_2R} = \left|h_{S_2R}\right|^2$.

The CDF and PDF of g_{S_2R} are respectively given by:

$$F_{g_{S_2R}}(x) = 1 - \exp\left(-\lambda_{S_2R}x\right), f_{g_{S_2R}}(x) = \lambda_{S_2R}\exp\left(-\lambda_{S_2R}x\right), \tag{7}$$

where $\lambda_{S_2R} = \left(d_{S_2R}\right)^{\beta}$, with d_{S_2R} denoting the distance between S_2 and R.

Similarly, Eq. (7) gives the marginal distribution of the desired channel power gain g_{S_2R}, rather than the CDF of the SINR in Eq. (6). The aggregate CCI term in Eq. (6) remains explicitly included and is treated through the same probability-averaging procedure used for the first source-to-relay link.

Using the instantaneous SINR in Eq. 6, the channel capacity of the $S_2 \to R$ link in the second time slot is given by:

$$C_{S_2R} = \frac{1}{3}\log_2\left(1+\varphi_{S_2R}\right) = \frac{1}{3}\log_2\left(1+\frac{\Psi_S g_{S_2R}}{\sum_{k=1}^{K}\Psi_I g_{I_kR}+1}\right). \quad (8)$$

Note: The relay R successfully decodes packet X_1 if $C_{S_1R} \geq C_{\text{th}}$ and it successfully decodes packet X_2 if $C_{S_2R} \geq C_{\text{th}}$. Otherwise, the corresponding packet-decoding operation is considered unsuccessful; i.e., when the corresponding instantaneous capacity is below C_{th}.

2.1.3 Third Time Slot

Therefore, four cases arise when R attempts to decode the packet-level payloads X_1 and X_2:

- Case 1: R successfully decodes both X_1 and X_2
  $C_{S_1R} \geq C_{\text{th}}$ and $C_{S_2R} \geq C_{\text{th}}$.
- Case 2: R successfully decodes X_1, but fails to decode X_2
  $C_{S_1R} \geq C_{\text{th}}$ and $C_{S_2R} < C_{\text{th}}$.
- Case 3: R successfully decodes X_2, but fails to decode X_1
  $C_{S_1R} < C_{\text{th}}$ and $C_{S_2R} \geq C_{\text{th}}$.
- Case 4: R fails to decode both X_1 and X_2
  $C_{S_1R} < C_{\text{th}}$ and $C_{S_2R} < C_{\text{th}}$.

a) Case 1: $C_{S_1R} \geq C_{\text{th}}$ and $C_{S_2R} \geq C_{\text{th}}$.

In this case, R successfully decodes both packet-level payloads X_1 and X_2, forms $X_3 = X_1 \oplus X_2$ and broadcasts the corresponding modulated symbol $x_3 = \mathcal{M}(X_3)$. The received signals at S_1 and S_2 are respectively expressed as:

$$\begin{aligned} y_{R\to S_1} &= \sqrt{P_S}h_{RS_1}x_3 + \sum_{k=1}^{K}\sqrt{P_I}h_{I_kS_1}u_k + n_{S_1}, \\ y_{R\to S_2} &= \sqrt{P_S}h_{RS_2}x_3 + \sum_{k=1}^{K}\sqrt{P_I}h_{I_kS_2}u_k + n_{S_2}, \end{aligned} \quad (9)$$

where h_{RS_1} and h_{RS_2} denote the channel coefficients of the $R \to S_1$ and $R \to S_2$ links, respectively.

The summation term in each received signal represents the aggregate CCI observed at the corresponding terminal. Therefore, the interference contributions are explicitly retained in the denominators of the instantaneous SINRs in Eqs. (10 and 11).

From the received signals in Eq. (9), the instantaneous SINRs at S_1 and S_2 are respectively obtained as:

$$\varphi_{RS_1} = \frac{P_S\left|h_{RS_1}\right|^2}{\sum_{k=1}^{K}P_I\left|h_{I_kS_1}\right|^2+\sigma_0^2} = \frac{\Psi_S g_{RS_1}}{\sum_{k=1}^{K}\Psi_I g_{I_kS_1}+1}, \quad (10)$$

$$\varphi_{RS_2} = \frac{P_S\left|h_{RS_2}\right|^2}{\sum_{k=1}^{K}P_I\left|h_{I_kS_2}\right|^2+\sigma_0^2} = \frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K}\Psi_I g_{I_kS_2}+1}, \quad (11)$$

where $g_{RS_1} = \left|h_{RS_1}\right|^2, g_{RS_2} = \left|h_{RS_2}\right|^2, g_{I_kS_1} = \left|h_{I_kS_1}\right|^2$ and $g_{I_kS_2} = \left|h_{I_kS_2}\right|^2$.

Note: Consistent with the reciprocity assumptions stated above, g_{S_iR} and g_{RS_i} are independent small-scale fading realizations, because they correspond to different time slots, but they share the same exponential rate parameter $\lambda_{S_iR}, i \in \{1,2\}$.

Under the exponential channel-power model stated above, the marginal CDFs and PDFs of $g_{I_kS_1}$ and $g_{I_kS_2}$ are given by:

$$\begin{aligned} F_{g_{I_kS_1}}(x) &= 1-\exp\left(-\lambda_{I_kS_1}x\right), F_{g_{I_kS_2}}(x) = 1-\exp\left(-\lambda_{I_kS_2}x\right), \\ f_{g_{I_kS_1}}(x) &= \lambda_{I_kS_1}\exp\left(-\lambda_{I_kS_1}x\right), f_{g_{I_kS_2}}(x) = \lambda_{I_kS_2}\exp\left(-\lambda_{I_kS_2}x\right), \end{aligned} \quad (12)$$

where $\lambda_{I_k S_1} = \left(d_{I_k S_1}\right)^{\beta}$ and $\lambda_{I_k S_2} = \left(d_{I_k S_2}\right)^{\beta}$, with $d_{I_k S_1}$ and $d_{I_k S_2}$ denoting the distances from interferer I_k to the source nodes.

Eq. (12) gives the marginal distributions of the individual interference-channel power gains. It does not represent the distribution of the aggregate CCI or of the terminal SINRs. The aggregate interference powers are explicitly included in Eqs. (10 and 11), respectively and are subsequently incorporated into the packet-delivery probability analysis by averaging over the K independent interference gains.

$$\sum_{k=1}^{K} \Psi_I g_{I_k S_1} \text{ and } \sum_{k=1}^{K} \Psi_I g_{I_k S_2},$$

The channel capacity for the links is calculated using the following formulas:

$$C_{RS_1} = \frac{1}{3}\log_2\left(1+\varphi_{RS_1}\right) = \frac{1}{3}\log_2\left(1+\frac{\Psi_S g_{RS_1}}{\sum_{k=1}^{K} \Psi_I g_{I_k S_1}+1}\right),$$

$$C_{RS_2} = \frac{1}{3}\log_2\left(1+\varphi_{RS_2}\right) = \frac{1}{3}\log_2\left(1+\frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K} \Psi_I g_{I_k S_2}+1}\right) \quad (13)$$

Note: If $C_{RS_1} \geq C_{\text{th}}$, then S_1 successfully decodes packet X_3 from the received modulated symbol $x_3 = \mathcal{M}(X_3)$; otherwise, the decoding operation fails. Similarly, S_2 successfully decodes packet X_3 if $C_{RS_2} \geq C_{\text{th}}$. After decoding packet X_3, terminal S_1 recovers $X_2 = X_3 \oplus X_1$, whereas S_2 recovers $X_1 = X_3 \oplus X_2$. Each terminal then stores the recovered fountain-coded packet for subsequent source-message decoding.

b) Case 2: $C_{S_1 R} \geq C_{\text{th}}$ and $C_{S_2 R} < C_{\text{th}}$.

In this case, R successfully decodes packet X_1, but fails to decode packet X_2. Therefore, it forwards X_1 to S_2 by transmitting the corresponding symbol $x_1 = \mathcal{M}(X_1)$ in the third time slot. The received signal at S_2, together with its SINR and channel capacity, is given as follows [7]:

$$y_{R\to S_2} = \sqrt{P_S} h_{RS_2} x_1 + \sum_{k=1}^{K} \sqrt{P_I} h_{I_k S_2} u_k + n_{S_2}, \quad (14)$$

$$\varphi_{RS_2} = \frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K} \Psi_I g_{I_k S_2}+1}, \quad (15)$$

$$C_{RS_2} = \frac{1}{3}\log_2\left(1+\varphi_{RS_2}\right) = \frac{1}{3}\log_2\left(1+\frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K} \Psi_I g_{I_k S_2}+1}\right). \quad (16)$$

c) Case 3: $C_{S_1 R} < C_{\text{th}}$ and $C_{S_2 R} \geq C_{\text{th}}$.

In this case, R successfully decodes packet X_2, but fails to decode packet X_1. Therefore, it forwards X_2 to S_1 by transmitting the corresponding symbol $x_2 = \mathcal{M}(X_2)$ in the third time slot. The received signal at S_1, together with its SINR and channel capacity, is given as follows [7]:

$$y_{R\to S_1} = \sqrt{P_S} h_{RS_1} x_2 + \sum_{k=1}^{K} \sqrt{P_I} h_{I_k S_1} u_k + n_{S_1}, \quad (17)$$

$$\varphi_{RS_1} = \frac{\Psi_S g_{RS_1}}{\sum_{k=1}^{K} \Psi_I g_{I_k S_1}+1}. \quad (18)$$

$$C_{RS_1} = \frac{1}{3}\log_2\left(1+\varphi_{RS_1}\right) = \frac{1}{3}\log_2\left(1+\frac{\Psi_S g_{RS_1}}{\sum_{k=1}^{K} \Psi_I g_{I_k S_1}+1}\right). \quad (19)$$

d) Case 4: $C_{S_1 R} < C_{\text{th}}$ and $C_{S_2 R} < C_{\text{th}}$.

In this case, if R fails to decode both X_1 and X_2, no data is transmitted in the third time slot. Consequently, both source nodes will not receive any new information during this phase, increasing the likelihood of an outage.

2.2 Probability Analysis Using the SM-3P Model

Because the total duration of one SM-3P packet-exchange round is normalized to one time unit and consists of three equal-duration transmission phases, the successful-decoding condition for an instantaneous SINR γ is transformed as:

$$\frac{1}{3}\log_2(1+\gamma) \geq C_{\text{th}} \Leftrightarrow \log_2(1+\gamma) \geq 3C_{\text{th}} \Leftrightarrow \gamma \geq 2^{3C_{\text{th}}} - 1 \triangleq \rho_{\text{th}}.$$

Thus, the factor three associated with the SM-3P transmission duration is absorbed into the SINR threshold ρ_{th} and is used throughout the following outage-probability derivation.

Because all channel power gains are continuously distributed, the probability of the boundary event $C = C_{\text{th}}$ is zero. Therefore, replacing the strict inequality by the non-strict inequality does not alter the derived outage probabilities. Nevertheless, the successful-decoding condition is consistently written using $\geq C_{\text{th}}$, whereas decoding failure is defined using $< C_{\text{th}}$. First, we derive the probability that packet X_1 is successfully delivered from S_1 to S_2. For S_2 to successfully receive packet X_1, both the $S_1 \to R$ and $R \to S_2$ transmissions must be successful. Since these transmissions occur in different time slots and their corresponding fading realizations are modeled as independent, the end-to-end success probability can be factorized as follows:

$$\theta_{X_1}^{\text{SM-3P}} = \Pr\left(C_{S_1R} \geq C_{\text{th}}\right) \Pr\left(C_{RS_2} \geq C_{\text{th}}\right)$$

$$= \Pr\left(\frac{1}{3}\log_2\left(1 + \frac{\Psi_S g_{S_1R}}{\sum_{k=1}^{K} \Psi_I g_{I_kR} + 1}\right) \geq C_{\text{th}}\right) \Pr\left(\frac{1}{3}\log_2\left(1 + \frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K} \Psi_I g_{I_kS_2} + 1}\right) \geq C_{\text{th}}\right)$$

$$= \Pr\left(\frac{\Psi_S g_{S_1R}}{\sum_{k=1}^{K} \Psi_I g_{I_kR} + 1} \geq \rho_{\text{th}}\right) \Pr\left(\frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K} \Psi_I g_{I_kS_2} + 1} \geq \rho_{\text{th}}\right), \tag{20}$$

Next, we calculate the probability term in Eq. (20) as follows:

$$\Pr\left(\frac{\Psi_S g_{S_1R}}{\sum_{k=1}^{K} \Psi_I g_{I_kR} + 1} \geq \rho_{\text{th}}\right) = \Pr\left(\Psi_S g_{S_1R} \geq \rho_{\text{th}} \Psi_I \sum_{k=1}^{K} g_{I_kR} + \rho_{\text{th}}\right) \tag{21}$$

$$= \Pr\left(g_{S_1R \geq \omega_1} \sum_{k=1}^{K} g_{I_kR} + \omega_2\right), \tag{21}$$

with

$$\omega_1 = \frac{\rho_{\text{th}} \Psi_I}{\Psi_S}, \omega_2 = \frac{\rho_{\text{th}}}{\Psi_S}.$$

It is worth emphasizing that the above analytical expressions do not require equal desired-signal and interference powers. The unequal-power case is directly captured through $\Psi_S = P_S/\sigma_0^2$ and $\Psi_I = P_I/\sigma_0^2$. In particular, the outage behavior is governed by both the desired-node SNR Ψ_S and the power ratio Ψ_I/Ψ_S.

When P_S and P_I increase proportionally, Ψ_I/Ψ_S remains constant. Consequently, ω_1 remains non-zero while ω_2 approaches zero and the system converges to an interference-limited outage floor. In contrast, when P_S increases while P_I remains fixed, both ω_1 and ω_2 approach zero. Therefore, the outage probability continues to decrease and does not approach the same interference-limited floor. Conversely, increasing P_I for a fixed P_S increases ω_1 and degrades the packet- and message-recovery probabilities.

These observations indicate that separate power allocation for the desired nodes and interference management for the co-channel interferers can significantly affect system performance. Increasing the transmit power of the sources and relay is effective when the aggregate interference power remains bounded. However, proportional power scaling at all transmitting nodes cannot eliminate the outage floor. Hence, desired-node power allocation should be combined with interference-aware scheduling, resource allocation, interference cancellation, admission control or constraints on the admissible interference power. When the interferers are uncoordinated, the system may instead adapt P_S according to the measured interference level. Such mechanisms are particularly relevant to energy-

constrained WSN deployments, where increasing P_S without considering the associated energy cost may not be practical. To evaluate the probability term in Eq. (21), we condition on the K independent interference-channel power gains by setting $g_{I_kR} = z_k, k = 1, \dots, K$.

Since one integration variable is associated with each co-channel interferer, averaging over their joint distribution requires a K-fold integral. Moreover, owing to the independence assumption, their joint PDF factorizes as: $f_{g_{I_1R},\dots,g_{I_KR}}(z_1, \dots, z_K) = \prod_{k=1}^{K} f_{g_{I_kR}}(z_k)$.

Accordingly, the probability term is evaluated as:

$$\Pr\left(g_{S_1R} \geq \omega_1 \sum_{k=1}^{K} g_{I_kR} + \omega_2\right) = \underbrace{\int_0^\infty \cdots \int_0^\infty}_{K-\text{fold integral}} \left[1 - F_{g_{S_1R}}\left(\omega_1 \sum_{k=1}^{K} z_k + \omega_2\right)\right] \prod_{k=1}^{K} f_{g_{I_kR}}(z_k) dz_1 \cdots dz_K$$

$$= \exp\left(-\lambda_{S_1R}\omega_2\right) \prod_{k=1}^{K} \int_0^\infty \lambda_{I_kR} \exp\left[-\left(\lambda_{I_kR} + \lambda_{S_1R}\omega_1\right)z_k\right] dz_k =$$

$$\exp\left(-\lambda_{S_1R}\omega_2\right) \prod_{k=1}^{K} \frac{\lambda_{I_kR}}{\lambda_{I_kR} + \lambda_{S_1R}\omega_1} \tag{22}$$

Using the same conditioning and K-fold averaging procedure, the success probability of the $R \to S_2$ link is obtained analogously. Multiplying the two independent link-success probabilities yields the per-round delivery probability of packet X_1 under the SM-3P protocol as:

$$\theta_{X_1}^{\text{SM}-3\text{P}} = \Pr\left(C_{S_1R} \geq C_{\text{th}}\right)\Pr\left(C_{RS_2} \geq C_{\text{th}}\right)$$

$$= \exp\left(-\left(\lambda_{S_1R} + \lambda_{S_2R}\right)\omega_2\right) \prod_{k=1}^{K} \frac{\lambda_{I_kR}\lambda_{I_kS_2}}{\left(\lambda_{I_kR} + \lambda_{S_1R}\omega_1\right)\left(\lambda_{I_kS_2} + \lambda_{S_2R}\omega_1\right)} \tag{23}$$

Therefore, the per-round delivery-failure probability of packet X_1 at S_2 is given by:

$$1 - \theta_{X_1}^{\text{SM}-3\text{P}} = 1 - \exp\left(-\left(\lambda_{S_1R} + \lambda_{S_2R}\right)\omega_2\right) \prod_{k=1}^{K} \frac{\lambda_{I_kR}\lambda_{I_kS_2}}{\left(\lambda_{I_kR} + \lambda_{S_1R}\omega_1\right)\left(\lambda_{I_kS_2} + \lambda_{S_2R}\omega_1\right)}. \tag{24}$$

Similarly, the per-round probability that S_1 successfully receives packet X_2 is given by:

$$\theta_{X_2}^{\text{SM}-3\text{P}} = \Pr\left(C_{S_2R} \geq C_{\text{th}}\right)\Pr\left(C_{RS_1} \geq C_{\text{th}}\right)$$

$$= \Pr\left(\frac{\Psi_S g_{S_2R}}{\sum_{k=1}^{K} \Psi_I g_{I_kR} + 1} \geq \rho_{\text{th}}\right) \Pr\left(\frac{\Psi_S g_{RS_1}}{\sum_{k=1}^{K} \Psi_I g_{I_kS_1} + 1} \geq \rho_{\text{th}}\right)$$

$$= \exp\left(-\left(\lambda_{S_1R} + \lambda_{S_2R}\right)\omega_2\right) \prod_{k=1}^{K} \frac{\lambda_{I_kR}\lambda_{I_kS_1}}{\left(\lambda_{I_kR} + \lambda_{S_2R}\omega_1\right)\left(\lambda_{I_kS_1} + \lambda_{S_1R}\omega_1\right)}. \tag{25}$$

The corresponding per-round delivery-failure probability of packet X_2 at S_1 is given by:

$$1 - \theta_{X_2}^{\text{SM}-3\text{P}} = 1 - \exp\left(-\left(\lambda_{S_1R} + \lambda_{S_2R}\right)\omega_2\right) \prod_{k=1}^{K} \frac{\lambda_{I_kR}\lambda_{I_kS_1}}{\left(\lambda_{I_kR} + \lambda_{S_2R}\omega_1\right)\left(\lambda_{I_kS_1} + \lambda_{S_1R}\omega_1\right)}. \tag{26}$$

Now, we consider the outage probabilities at S_1 and S_2. The desired and interfering channel realizations are assumed to vary independently from one packet-exchange round to another, while the node locations, transmit powers and channel statistics remain unchanged. Therefore, the packet-delivery outcomes over the Q exchange rounds are modeled as independent trials with the same per-round success probability.

Let N_2 denote the number of fountain-coded packets generated by S_1 and successfully received by S_2 over Q exchange rounds. Since S_2 requires at least H successfully received fountain-coded packets to recover the source message, an outage occurs at S_2 when $N_2 < H$; i.e., $N_2 = 0,1, \dots, H-1$. Therefore, the outage probability at S_2 is expressed as follows:

$$\text{OP}_{S_2}^{\text{SM}-3\text{P}} = \sum_{N_2=0}^{H-1} \binom{Q}{N_2}, \left(\theta_{X_1}^{\text{SM}-3\text{P}}\right)^{N_2} \left(1 - \theta_{X_1}^{\text{SM}-3\text{P}}\right)^{Q-N_2}$$

$$= \sum_{N_2=0}^{H-1} \binom{Q}{N_2} \left(\exp\left(-\left(\lambda_{S_1R} + \lambda_{S_2R}\right)\omega_2\right) \prod_{k=1}^{K} \frac{\lambda_{I_kR}\lambda_{I_kS_2}}{\left(\lambda_{I_kR} + \lambda_{S_1R}\omega_1\right)\left(\lambda_{I_kS_2} + \lambda_{S_2R}\omega_1\right)}\right)^{N_2}$$

$$\times\left(1-\exp\left(-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\omega_2\right)\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_2}}{\left(\lambda_{I_kR}+\lambda_{S_1R}\omega_1\right)\left(\lambda_{I_kS_2}+\lambda_{S_2R}\omega_1\right)}\right)^{Q-N_2}. \quad (27)$$

where $\binom{Q}{N_i}$ denotes the binomial coefficient.

Similarly, let N_1 denote the number of fountain-coded packets generated by S_2 and successfully received by S_1 over Q exchange rounds. An outage occurs at S_1 if $N_1 < H$; i.e., $N_1 = 0,1,2,\ldots,H-1$. Therefore, the outage probability at S_1 is given by:

$$\mathrm{OP}_{S_1}^{\mathrm{SM-3P}} = \sum_{N_1=0}^{H-1}\binom{Q}{N_1}\left(\theta_{X_2}^{\mathrm{SM-3P}}\right)^{N_1}\left(1-\theta_{X_2}^{\mathrm{SM-3P}}\right)^{Q-N_1}$$

$$= \sum_{N_1=0}^{H-1}\binom{Q}{N_1}\left(\exp\left[-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\omega_2\right]\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_1}}{\left(\lambda_{I_kR}+\lambda_{S_2R}\omega_1\right)\left(\lambda_{I_kS_1}+\lambda_{S_1R}\omega_1\right)}\right)^{N_1}$$

$$= \sum_{N_1=0}^{H-1}\binom{Q}{N_1}\left(1-\exp\left[-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\omega_2\right]\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_1}}{\left(\lambda_{I_kR}+\lambda_{S_2R}\omega_1\right)\left(\lambda_{I_kS_1}+\lambda_{S_1R}\omega_1\right)}\right)^{Q-N_1} \quad (28)$$

2.3 TWR Model with Four Phases (SM-4P)

Similarly, because one normalized SM-4P packet-exchange round consists of four equal-duration transmission phases, the successful-decoding condition is transformed as:

$$\frac{1}{4}\log_2(1+\gamma) \geq C_{\mathrm{th}} \Leftrightarrow \log_2(1+\gamma) \geq 4C_{\mathrm{th}} \Leftrightarrow \gamma \geq 2^{4C_{\mathrm{th}}} - 1 \triangleq \xi_{\mathrm{th}}.$$

The same non-strict successful-decoding convention adopted in Sub-section 2.2 is used for the SM-4P analysis. Hence, the factor four is incorporated into the SM-4P SINR threshold ξ_{th} used in the subsequent analysis.

The SM-4P-FC benchmark employs the same network topology, node locations, transmit-power definitions, CCI model, channel distributions and independent block-fading assumptions introduced in Sub-section 2.1. Unlike the two SM-3P schemes, SM-4P-FC does not perform a packet-level XOR operation at the relay. Instead, the two communication directions are forwarded separately over four transmission phases.

The corresponding four-phase transmission model is depicted in Figure 2. In the first phase, S_1 transmits its packet to R. In the second phase, R forwards the recovered packet to S_2. In the third phase, S_2 transmits its packet to R and in the fourth phase, R forwards the recovered packet to S_1. Under the common channel and interference model, the instantaneous link capacities are given by:

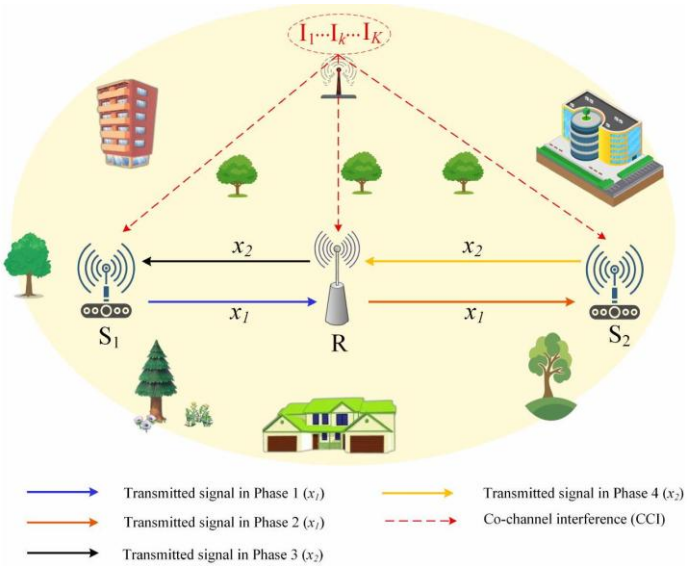


Figure 2. Four-phase TWR model operating in a co-channel interference environment.

$$C_{S_1R}=\frac{1}{4}\log_2\left(1+\frac{\Psi_S g_{S_1R}}{\sum_{k=1}^{K}\Psi_I g_{I_kR}+1}\right),\quad C_{RS_2}=\frac{1}{4}\log_2\left(1+\frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K}\Psi_I g_{I_kS_2}+1}\right),$$
$$C_{S_2R}=\frac{1}{4}\log_2\left(1+\frac{\Psi_S g_{S_2R}}{\sum_{k=1}^{K}\Psi_I g_{I_kR}+1}\right),\quad C_{RS_1}=\frac{1}{4}\log_2\left(1+\frac{\Psi_S g_{RS_1}}{\sum_{k=1}^{K}\Psi_I g_{I_kS_1}+1}\right). \tag{29}$$

Using C_{S_1R} and C_{RS_2} in Eq. (29), the per-round probability that S_2 successfully receives the encoded packet X_1 is calculated as follows:

$$\theta_{X_1}^{\mathrm{SM-4P}}=\Pr\left(\frac{1}{4}\log_2\left(1+\frac{\Psi_S g_{S_1R}}{\sum_{k=1}^{K}\Psi_I g_{I_kR}+1}\right)\geq C_{\mathrm{th}}\right)\Pr\left(\frac{1}{4}\log_2\left(1+\frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K}\Psi_I g_{I_kS_2}+1}\right)\geq C_{\mathrm{th}}\right)$$
$$=\Pr\left(\frac{\Psi_S g_{S_1R}}{\sum_{k=1}^{K}\Psi_I g_{I_kR}+1}\geq \xi_{\mathrm{th}}\right)\Pr\left(\frac{\Psi_S g_{RS_2}}{\sum_{k=1}^{K}\Psi_I g_{I_kS_2}+1}\geq \xi_{\mathrm{th}}\right)$$
$$=\exp\left[-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\vartheta_2\right]\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_2}}{\left(\lambda_{I_kR}+\lambda_{S_1R}\vartheta_1\right)\left(\lambda_{I_kS_2}+\lambda_{S_2R}\vartheta_1\right)}. \tag{30}$$

where $\vartheta_1=\xi_{\mathrm{th}}\Psi_I/\Psi_S$ and $\vartheta_2=\xi_{\mathrm{th}}/\Psi_S$.

Next, the per-round probability that S_1 successfully receives packet X_2 is calculated as:

$$\theta_{X_2}^{\mathrm{SM-4P}}=\Pr\left(\frac{1}{4}\log_2\left(1+\frac{\Psi_S g_{S_2R}}{\sum_{k=1}^{K}\Psi_I g_{I_kR}+1}\right)\geq C_{\mathrm{th}}\right)\times\Pr\left(\frac{1}{4}\log_2\left(1+\frac{\Psi_S g_{RS_1}}{\sum_{k=1}^{K}\Psi_I g_{I_kS_1}+1}\right)\geq C_{\mathrm{th}}\right)$$
$$=\Pr\left(\frac{\Psi_S g_{S_2R}}{\sum_{k=1}^{K}\Psi_I g_{I_kR}+1}\geq \xi_{\mathrm{th}}\right)\Pr\left(\frac{\Psi_S g_{RS_1}}{\sum_{k=1}^{K}\Psi_I g_{I_kS_1}+1}\geq \xi_{\mathrm{th}}\right)$$
$$=\exp\left[-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\vartheta_2\right]\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_1}}{\left(\lambda_{I_kR}+\lambda_{S_2R}\vartheta_1\right)\left(\lambda_{I_kS_1}+\lambda_{S_1R}\vartheta_1\right)}. \tag{31}$$

Finally, the outage probabilities at S_1 and S_2 under the SM-4P protocol are calculated as follows:

$$\mathrm{OP}_{S_1}^{\mathrm{SM-4P}}=\sum_{N_1=0}^{H-1}\binom{Q}{N_1}\left(\theta_{X_2}^{\mathrm{SM-4P}}\right)^{N_1}\left(1-\theta_{X_2}^{\mathrm{SM-4P}}\right)^{Q-N_1}$$
$$=\sum_{N_1=0}^{H-1}\binom{Q}{N_1}\left(\exp\left[-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\vartheta_2\right]\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_1}}{\left(\lambda_{I_kR}+\lambda_{S_2R}\vartheta_1\right)\left(\lambda_{I_kS_1}+\lambda_{S_1R}\vartheta_1\right)}\right)^{N_1}$$
$$\times\left(1-\exp\left[-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\vartheta_2\right]\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_1}}{\left(\lambda_{I_kR}+\lambda_{S_2R}\vartheta_1\right)\left(\lambda_{I_kS_1}+\lambda_{S_1R}\vartheta_1\right)}\right)^{Q-N_1}. \tag{32}$$

$$\mathrm{OP}_{S_2}^{\mathrm{SM-4P}}=\sum_{N_2=0}^{H-1}\binom{Q}{N_2}\left(\theta_{X_1}^{\mathrm{SM-4P}}\right)^{N_2}\left(1-\theta_{X_1}^{\mathrm{SM-4P}}\right)^{Q-N_2}$$
$$=\sum_{N_2=0}^{H-1}\binom{Q}{N_2}\left(\exp\left[-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\vartheta_2\right]\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_2}}{\left(\lambda_{I_kR}+\lambda_{S_1R}\vartheta_1\right)\left(\lambda_{I_kS_2}+\lambda_{S_2R}\vartheta_1\right)}\right)^{N_2}$$
$$\times\left(1-\exp\left[-\left(\lambda_{S_1R}+\lambda_{S_2R}\right)\vartheta_2\right]\prod_{k=1}^{K}\frac{\lambda_{I_kR}\lambda_{I_kS_2}}{\left(\lambda_{I_kR}+\lambda_{S_1R}\vartheta_1\right)\left(\lambda_{I_kS_2}+\lambda_{S_2R}\vartheta_1\right)}\right)^{Q-N_2}. \tag{33}$$

2.4 Sum Throughput Analysis

To complement the outage-probability analysis, we further evaluate the bandwidth-normalized throughput of the two-way relay system. Under the adopted fixed- Q, no-early-stopping packet-exchange policy, each source message contains H useful source packets and all Q packet-exchange rounds are used, even if a destination could satisfy its recovery condition before the final round. Therefore, H/Q represents the ratio between the number of useful source packets and the fixed total

number of packet-exchange rounds.

For protocol $p \in \{\mathrm{SM}-3\mathrm{P}, \mathrm{SM}-4\mathrm{P}\}$, the throughput at terminal $S_i, i \in \{1,2\}$, is defined as:

$$\eta_{S_i}^{(p)} = \frac{H}{Q} C_{\mathrm{th}} \left(1 - \mathrm{OP}_{S_i}^{(p)}\right), \tag{34}$$

where $1 - \mathrm{OP}_{S_i}^{(p)}$ denotes the probability that terminal S_i successfully recovers the source message. Accordingly, the bidirectional sum throughput is given by:

$$\eta_{\mathrm{sum}}^{(p)} = \eta_{S_1}^{(p)} + \eta_{S_2}^{(p)} = \frac{H}{Q} C_{\mathrm{th}} \left[2 - \mathrm{OP}_{S_1}^{(p)} - \mathrm{OP}_{S_2}^{(p)}\right]. \tag{35}$$

This definition applies to the fixed- Q, no-early-stopping policy adopted for the SM-3P-FC, SM-4P-FC and SM-3P-NoFC comparisons. If early stopping were employed, the denominator would instead depend on the average number of packet-exchange rounds required for successful recovery and would require a separate analysis.

The throughput is normalized by the system bandwidth and is therefore expressed in bit/s/Hz. No additional factor of 1/3 or 1/4 is introduced in Eqs. (34 and 35), because the number of transmission phases has already been incorporated into the instantaneous capacities and the corresponding outage probabilities of the SM-3P and SM-4P protocols.

The above metric follows the ideal packet-level fountain-coding abstraction adopted in this work. Consequently, packet-header overhead, additional encoded packets required for practical decoding, decoding complexity, processing delay, memory consumption and the corresponding energy cost are not included in the throughput calculation. Therefore, the reported values represent idealized bandwidth-normalized throughput.

At the present abstraction level, practical fountain-code overhead may be approximated by evaluating the outage probabilities appearing in Eqs. (34 and 35) using an effective recovery threshold $H_{\mathrm{eff}} > H$, while retaining H/Q as the useful-payload factor. This effective-threshold substitution does not represent the complete finite-length decoding behavior. A code-specific throughput analysis would additionally require the corresponding finite-length decoding-success function, together with explicit accounting for packet-header, coding, signaling, processing, memory and energy overhead.

2.5 SM-3P Uncoded Fixed-Repetition Benchmark

To isolate the packet-recovery gain provided by fountain coding, we consider a same-budget three-phase uncoded fixed-repetition benchmark, hereafter denoted by SM-3P-NoFC. The benchmark retains the same three-phase DNC physical transmission schedule, network topology, transmit-power definitions, CCI model, channel assumptions, source-block size H and total budget of Q packet-exchange rounds as SM-3P-FC. It differs only in its packet-generation and source-message-recovery rules.

Each source message consists of H distinct uncoded packets. These packets are transmitted according to a pre-determined fixed-repetition schedule over the same Q packet-exchange rounds used by SM-3P-FC. The fixed schedule requires neither packet-level feedback nor adaptive retransmission and therefore provides a controlled same-budget comparison. It is not intended to represent all possible non-fountain coding strategies, such as feedback-based ARQ or a separate fixed-rate channel code.

Let

$$L = \left\lfloor \frac{Q}{H} \right\rfloor, r = Q - LH, \tag{36}$$

where $0 \leq r < H$. For the considered benchmark, $Q \geq H$ is assumed, so that every uncoded source packet is transmitted at least once. Accordingly, r source packets are transmitted $L + 1$ times, whereas the remaining $H - r$ packets are transmitted L times. Thus, the Q transmission opportunities are distributed as evenly as possible among the H uncoded source packets.

The relay operation, DNC processing, channel model, transmit powers and interference conditions remain identical to those of the fountain-coded SM-3P protocol. Therefore, the per-round successful packet-delivery probabilities remain $\theta_{X_1}^{\mathrm{SM}-3\mathrm{P}}$ and $\theta_{X_2}^{\mathrm{SM}-3\mathrm{P}}$. However, without fountain coding,

successful message recovery requires every one of the H distinct source packets to be received at least once. Under the adopted independent block-fading model across packet-exchange rounds, the delivery outcomes of the repeated transmissions are mutually independent.

Consequently, the outage probability at S_2 is given by:

$$\mathrm{OP}_{S_2}^{\mathrm{SM-3P-NoFC}} = 1 - \left[1 - \left(1 - \theta_{X_1}^{\mathrm{SM-3P}}\right)^{L+1}\right]^{r} \times \left[1 - \left(1 - \theta_{X_1}^{\mathrm{SM-3P}}\right)^{L}\right]^{H-r}. \tag{37}$$

Similarly, the outage probability at S_1 is:

$$\mathrm{OP}_{S_1}^{\mathrm{SM-3P-NoFC}} = 1 - \left[1 - \left(1 - \theta_{X_2}^{\mathrm{SM-3P}}\right)^{L+1}\right]^{r} \times \left[1 - \left(1 - \theta_{X_2}^{\mathrm{SM-3P}}\right)^{L}\right]^{H-r}. \tag{38}$$

When $Q = H$, each uncoded packet is transmitted once and the expressions reduce to:

$$\mathrm{OP}_{S_2}^{\mathrm{SM-3P-NoFC}} = 1 - \left(\theta_{X_1}^{\mathrm{SM-3P}}\right)^{H},$$

and

$$\mathrm{OP}_{S_1}^{\mathrm{SM-3P-NoFC}} = 1 - \left(\theta_{X_2}^{\mathrm{SM-3P}}\right)^{H}.$$

The corresponding bidirectional sum throughput is:

$$\eta_{\mathrm{sum}}^{\mathrm{SM-3P-NoFC}} = \frac{H}{Q} C_{\mathrm{th}} [2 - \mathrm{OP}_{S_1}^{\mathrm{SM-3P-NoFC}} - \mathrm{OP}_{S_2}^{\mathrm{SM-3P-NoFC}}] \tag{39}$$

This benchmark uses the same source-block size, transmission budget, three-phase protocol and DNC operation as the fountain-coded SM-3P system. Hence, the comparison isolates the benefit of allowing any H successfully received encoded packets to recover the source message, rather than requiring all H distinct uncoded packets.

3. SIMULATION AND ANALYTICAL RESULTS

This section presents the analytical and MC simulation results for the OPs and bidirectional sum throughputs of the SM-3P-FC, SM-3P-NoFC and SM-4P-FC schemes. The analytical curves are evaluated using the closed-form outage-probability expressions and the throughput expressions derived in Section 2, whereas the simulation results are obtained by independently generating the channel realizations and evaluating the corresponding packet-recovery events.

For consistency with the analytical notation, the labels SM-3P and SM-4P used in Figures 3-8 refer to the fountain-coded SM-3P-FC and SM-4P-FC protocols, respectively.

For each plotted point, the MC results are obtained from $N_{\mathrm{MC}} = 10^6$ independently generated channel realizations. The pseudo-random-number generator is initialized using the fixed seed 2026 with the Mersenne Twister algorithm. In all figures, analytical results are represented by continuous curves, whereas MC results are represented by discrete markers.

Unless otherwise specified, Figures 3-10 use the proportional power setting $P_S = P_I = P$ or equivalently $\Psi_S = \Psi_I = \Psi$. This setting is adopted to investigate the interference-limited behavior that occurs when the desired and interfering powers increase at the same rate. It is not an analytical restriction, because the derived expressions allow arbitrary and unequal values of Ψ_S and Ψ_I.

Figure 11 explicitly relaxes this numerical assumption by varying Ψ_S under three interference conditions: no CCI, fixed $\Psi_I = 5$ dB and proportional $\Psi_I = \Psi_S$.

For the interference-free case $K = 0$, the standard conventions $\sum_{k=1}^{0}(\cdot) = 0$ and $\prod_{k=1}^{0}(\cdot) = 1$ are adopted. The close agreement between the analytical and simulation results supports the accuracy of the derived expressions. Unless otherwise specified, the system and simulation parameters are summarized in Table 3.

For the controlled equal-large-scale-interference benchmark, the K interferers are assigned the same nominal distance profile, generated from the reference coordinate (0.3, -1.5). This controlled arrangement isolates the effect of the number of interferers while maintaining the same average large-scale propagation condition for every interferer. It is not intended to represent a literal deployment in which multiple physical interferers occupy exactly the same location. The small-scale fading

coefficients associated with the K interferers are independently generated and hence their instantaneous interference contributions are independent.

Table 3. Simulation parameters and numerical settings for Figures 3-11.

Parameter	Meaning	Value or setting
S_1	Coordinate of source node S_1	(0,0)
S_2	Coordinate of source node S_2	(1,0)
R	Coordinate of relay node R	(x_R, 0), where $0 < x_R < 1$
I_k	Nominal reference coordinate for the controlled equal-large-scale interference benchmark	$(0.3, -1.5), k = 1, \dots, K$
β	Path-loss exponent	3
σ_0^2	AWGN noise power at each receiving node	1
λ_{ab}	Exponential rate parameter of the channel power gain on link $a \to b$	$\lambda_{ab} = d_{ab}^{\beta}$
Fading	Small-scale fading model for the desired and interfering links	Independent block-Rayleigh fading
Channel update	Variation of the channel realizations across transmissions	Constant within one time slot and independently regenerated across time slots and exchange rounds
P_S, P_I	Desired-node and co-channel-interferer transmit powers; arbitrary in the analytical framework	$P_S = P_I = P$ for Figs. 3-10; separately specified through Ψ_S and Ψ_I for Fig. 11.
Ψ	Common normalized transmit power under $P_S = P_I = P$, where $\Psi = P/\sigma_0^2$	0,3,6, ... ,30 dB for Figs. 3, 4 and 10; 5 dB for Figs. 5-9.
Ψ_S, Ψ_I	Separate normalized desired-node and co-channel-interference powers used for the CCI-power comparison	For Fig. 11: Ψ_S = 0,3,6, ... ,30 dB; no CCI: $K = 0$; fixed CCI: $K = 2$, $\Psi_I = 5$ dB; proportional CCI: $K = 2, \Psi_I = \Psi_S$
C_{th}	Target spectral-efficiency threshold	0.1 bit/s/Hz for Figs. 3-8 and 10-11; 0.1, 0.2, . . , , 1.0 bit/s/Hz for Fig. 9.
H	Source-block size and recovery threshold for the fountain-coded schemes	3 for Figs. 3, 5, 6 and 811; 4 for Fig. 4; 4, 5, ..., 10 for Fig. 7.
Q	Maximum number of packet-exchange rounds	4 for Fig. 3; 5 for Figs. 4, 6 and 8-11; 3,4,5,6 for Fig. 5; 10 for Fig. 7.
x_R	X-coordinate of relay R on the $S_1 - S_2$ line segment	0.4 for Fig. 3; 0.75 for Fig. 4; 0.35 for Figs. 5, 711; 0.1, 0.2, . . , 0.9 for Fig. 6.
K	Number of co-channel interferers	2 for Figs. 3-6, 9 and 10; 3 for Fig. 7; 1,2,...,10 for Fig. 8; 0 or 2 for Fig. 11.
N_{MC}	Number of independently generated channel realizations for each plotted point	10^6
Random seed	Initialization of the MATLAB pseudo-random-number generator	2026, Mersenne Twister
Software	Software used for the analytical evaluations and Monte Carlo simulations	MATLAB R2024b

The numerical evaluations are conducted in a two-dimensional Cartesian coordinate system, where $S_1 = (0,0), S_2 = (1,0)$ and $R = (x_R, 0)$. All K co-channel interferers are assigned the same nominal coordinate in the reported numerical experiments. This fixed-location setting is not intended to represent a general spatial deployment. Rather, it provides a controlled conditional benchmark that isolates the impact of the number of interferers under identical large-scale propagation conditions.

With $S_1 = (0,0), S_2 = (1,0), R = (x_R, 0)$ and $I_k = (0.3, -1.5)$, the distances from each co-channel interferer to the receiving nodes are explicitly determined as: $d_{I_k S_1} = \sqrt{0.3^2 + 1.5^2}$, $d_{I_k S_2} = \sqrt{(1 - 0.3)^2 + 1.5^2}$ and $d_{I_k R} = \sqrt{(x_R - 0.3)^2 + 1.5^2}, k = 1, \dots, K$.

Accordingly, the corresponding exponential rate parameters used in the analytical and MC evaluations

are $\lambda_{I_k S_1} = d_{I_k S_1}^{\beta}, \lambda_{I_k S_2} = d_{I_k S_2}^{\beta}, \lambda_{I_k R} = d_{I_k R}^{\beta}$.

Although all K interferers are assigned the same nominal coordinate in the controlled numerical benchmark, their small-scale fading coefficients are independently generated. Thus, they share the same large-scale channel parameters, but produce independent instantaneous interference contributions.

A more general deployment can incorporate randomly distributed interferers through a spatial point-process model [30]. For a fixed number K, the interferer locations may be modeled by a binomial point process within a bounded deployment region. Alternatively, a Poisson point process may be employed when both the number and locations of the active interferers within a bounded observation region are random.

Conditioned on a spatial realization Φ_I, the present link-level expressions remain applicable by assigning each interferer location $\mathbf{x}_{I_k} \in \Phi_I$ and setting $d_{I_k b} = \|\mathbf{x}_{I_k} - \mathbf{x}_b\|, \lambda_{I_k b} = d_{I_k b}^{\beta}$, for each receiving node $b \in \{R, S_1, S_2\}$. Under the adopted exponential channel-power model, this definition gives $\mathbb{E}[g_{I_k b} \mid \Phi_I] = \frac{1}{\lambda_{I_k b}} = d_{I_k b}^{-\beta}$. A spatially averaged outage probability would then be obtained as $\overline{\mathrm{OP}}_{S_i} = \mathbb{E}_{\Phi_I}[\mathrm{OP}_{S_i} \mid \Phi_I]$, which requires an additional averaging operation over the random network geometry. Such a stochastic-geometry extension is beyond the scope of the present fixed- K, fixed-location analysis.

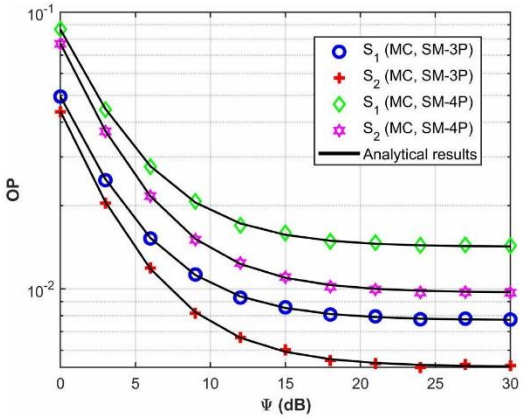


Figure 3. OP *versus* Ψ for $x_R = 0.4, H = 3, Q = 4$ and $K = 2$.

Figure 3 plots the OPs at S_1 and S_2 for the SM-3P and SM-4P protocols as functions of Ψ (in dB), with $H = 3, Q = 4$ and $K = 2$. As observed, the OP at both source nodes decreases as Ψ increases. This is because, in the low- and moderate-SNR regions, increasing Ψ improves the received signal quality and hence raises the successful packet decoding probability. However, since the interferers are also assumed to use the same transmit power scaling; i.e., $P_S = P_I = P$, the co-channel interference power increases together with the desired signal power. Therefore, at high Ψ, the system becomes interference-limited and the OP curves approach an outage floor. The figure also shows that the OPs achieved by the SM-3P scheme are always lower than those of the SM-4P scheme. This is because the three-phase protocol provides a larger pre-log factor (1/3 instead of 1/4), resulting in a higher instantaneous link capacity and, consequently, a lower outage probability. In addition, the OP at S_1 is higher than that at S_2, mainly due to the relay placement and the considered locations of the co-channel interferers.

Figure 4 plots the OPs at S_1 and S_2 for the SM-3P and SM-4P protocols as functions of Ψ, with $x_R = 0.75, H = 4, Q = 5$ and $K = 2$. Similar to Figure 3, the OPs initially decrease as Ψ increases and subsequently approach non-zero outage floors. This behavior occurs, because the desired nodes and co-channel interferers use the same power scaling; i.e., $P_S = P_I = P$, causing the system to become interference-limited at high Ψ.

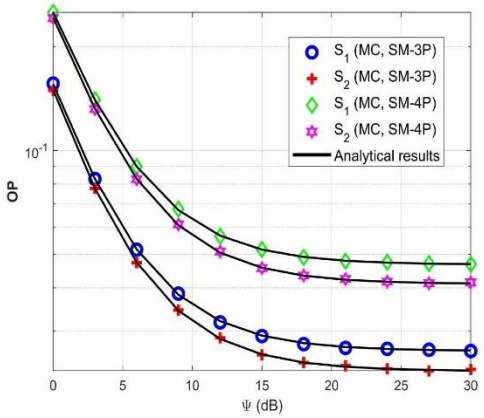


Figure 4. OP *versus* Ψ for $x_R = 0.75, H = 4, Q = 5$ and $K = 2$.

The SM-3P protocol consistently achieves lower OPs than the SM-4P protocol, because its larger pre-log factor results in a lower SINR threshold for the same target spectral efficiency. The difference between the OPs at S_1 and S_2 is primarily caused by the asymmetric relay and interferer geometry. Compared with Figure 3, relocating the relay to $x_R = 0.75$ changes the balance between the two end-to-end communication directions, whereas the different values of H and Q mainly affect the overall message-recovery probability.

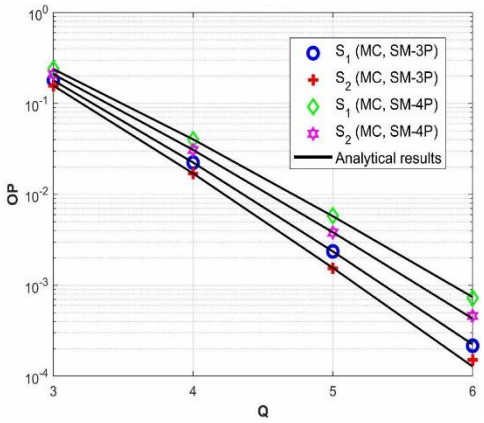


Figure 5. OP *versus* Q for $\Psi = 5$ dB, $x_R = 0.35, H = 3$ and $K = 2$.

Figure 5 plots the OPs at S_1 and S_2 for the SM-3P and SM-4P protocols as functions of the number of exchange rounds Q, with $\Psi = 5$ dB, $x_R = 0.35, H = 3$ and $K = 2$. It can be observed that the OPs at both source nodes under the SM-3P scheme remain lower than those under the SM-4P scheme. This is because the analyzed SM-3P protocol uses a larger pre-log factor than the four-phase protocol, resulting in a higher instantaneous link capacity and, consequently, a higher successful packet-decoding probability. The figure also shows that the OP at both S_1 and S_2 decreases as Q increases. This is because a larger Q increases the likelihood that the source nodes can successfully collect at least H fountain-coded packets required for successful message recovery, thereby reducing the outage probability. It is worth noting that the horizontal axis starts from $Q = 3$, because $H = 3$ is fixed in this experiment. Since at least H successfully received fountain-coded packets are required for message recovery, the cases with $Q < H$ are trivial, as successful recovery is impossible and the outage probability becomes unity. Therefore, only the meaningful region $Q \geq H$ is presented in the figure.

Figure 6 plots the OPs at S_1 and S_2 for the SM-3P and SM-4P protocols as functions of the relay location x_R, with $\Psi = 5$ dB, $Q = 5, H = 3$ and $K = 2$. It can be observed that the relay location has a pronounced impact on the outage performance of both source nodes. As x_R varies, the distances of the two hops change simultaneously, so improving one source-to-relay link may degrade the corresponding relay-to-source link on the other side. Therefore, a trade-off arises between the two transmission hops, which leads to a non-monotonic OP behavior. As a result, there exists an optimal relay region in which the OP of the source nodes is minimized.

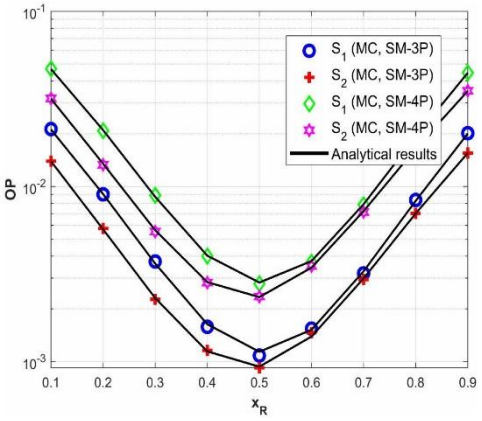


Figure 6. OP *versus* x_R for $\Psi = 5$ dB, $Q = 5, H = 3$ and $K = 2$.

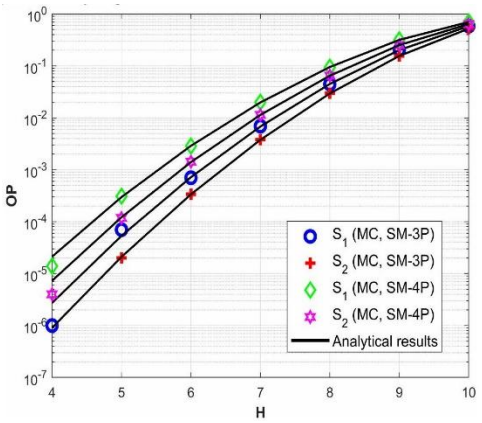


Figure 7. OP versus H for $\Psi = 5$ dB, $Q = 10, x_R = 0.35$ and $K = 3$.

Figure 7 shows the OPs at S_1 and S_2 as functions of the recovery threshold H. For a fixed $Q = 10$, the OPs increase as H increases, because a larger number of successfully received fountaincoded packets is required for source-message recovery. Therefore, the probability of collecting an insufficient number of packets becomes higher. The SM-3P protocol consistently achieves lower OPs than the SM-4P protocol over the entire considered range, further demonstrating the reliability advantage of the three-phase transmission structure.

The horizontal axis starts from $H = 4$, because $H = 3$ is used as the baseline recovery threshold in Figures 3, 5, 6 and 8-11, whereas Figure 7 focuses on stricter message-recovery requirements. All considered values satisfy $H \leq Q = 10$.

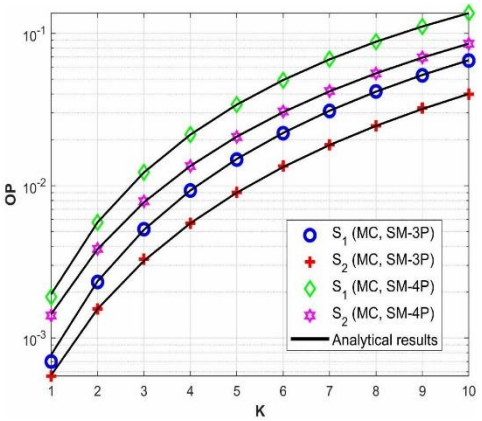


Figure 8. OP *versus* K for $\Psi = 5$ dB, $Q = 5, x_R = 0.35$ and $H = 3$.

Figure 8 plots the OPs at S_1 and S_2 for the SM-3P and SM-4P protocols as functions of K, with $\Psi = 5$ dB, $Q = 5, x_R = 0.35$ and $H = 3$. As observed, the OP at both nodes increases as K increases. This is because a larger number of co-channel interferers causes more severe interference on the transmission links, which reduces the received SINR and, consequently, the successful decoding

probability of fountain-coded packets. Therefore, the probability that the sources fail to collect the required H packets becomes higher, resulting in a larger OP. The figure also shows that the OPs achieved by the SM-3P scheme are always lower than those of the SM-4P scheme, confirming the effectiveness of the analyzed SM-3P protocol under stronger co-channel interference.

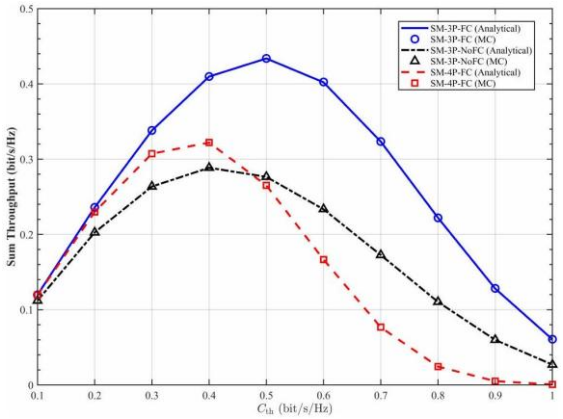


Figure 9. Bidirectional sum throughput *versus* C_{th} for the SM-3P-FC, SM-3P-NoFC and SM-4P-FC schemes with $\Psi = 5$ dB, $H = 3, Q = 5, x_R = 0.35$ and $K = 2$.

Figure 9 compares the bidirectional sum throughputs of the SM-3P-FC, SM-3P-NoFC and SM-4P-FC schemes as functions of C_{th}. Close agreement is observed between the analytical curves and the Monte Carlo markers over the evaluated C_{th} grid, supporting the accuracy of the derived throughput expressions, including the newly introduced SM-3P-NoFC benchmark. The maximum absolute deviations between the analytical and MC results over the evaluated C_{th} grid are $4.558 \times 10^{-4}, 2.345 \times 10^{-4}$ and 2.890×10^{-4} for SM-3P-FC, SM-3P-NoFC and SM-4P-FC, respectively. These maximum deviations occur at $C_{\text{th}} = 0.6, 0.6$ and 0.7 bit/s/Hz, respectively.

For all three schemes, the sum throughput initially increases with C_{th}, reaches a maximum and then decreases. At small C_{th}, the increase in spectral efficiency dominates the corresponding reliability loss. At larger C_{th}, however, the increased SINR threshold causes a rapid increase in outage probability and the reliability loss outweighs the spectral-efficiency improvement.

Under the common Figure 9 setting and over the evaluated target-spectral-efficiency grid $C_{\text{th}} \in \{0.1, 0.2, \ldots, 1.0\}$ bit /s/Hz, the analytical SM-3P-FC curve reaches its maximum evaluated sum throughput of approximately 0.433753 bit/s/Hz at $C_{\text{th}} = 0.5$ bit/s/Hz. By comparison, the analytical SM-3P-NoFC and SM-4P-FC curves reach their corresponding maxima of approximately 0.288574 and 0.322001 bit/s/Hz, respectively, at $C_{\text{th}} = 0.4$bit/s/Hz. Accordingly, within the considered numerical setting, the maximum analytical throughput of SM-3P-FC is approximately 50.3% higher than that of SM-3P-NoFC and approximately 34.7% higher than that of SM-4P-FC.

The SM-4P-FC scheme outperforms the SM-3P-NoFC benchmark at relatively small values of C_{th}, indicating that the recovery flexibility provided by fountain coding may compensate for the additional transmission phase in this region. However, for the sampled points satisfying $C_{\text{th}} \geq 0.5$ bit/s/Hz, SM-3P-NoFC outperforms SM-4P-FC, because the lower SINR threshold associated with the three-phase structure becomes increasingly important. These results demonstrate that both fountain coding and the reduction in the number of transmission phases contribute to the throughput improvement of the proposed SM-3P-FC scheme.

To further isolate the reliability contributions of fountain coding and the three-phase transmission structure, Figure 10 compares the OPs of the SM-3P-FC, SM-3P-NoFC and SM-4P-FC schemes.

The analytical curves are in close agreement with the MC markers for all three schemes at both terminals. The maximum absolute MC-analytical deviation among the six plotted cases is below 7.22×10^{-4}, further validating the newly introduced SM-3P-NoFC outage expressions.

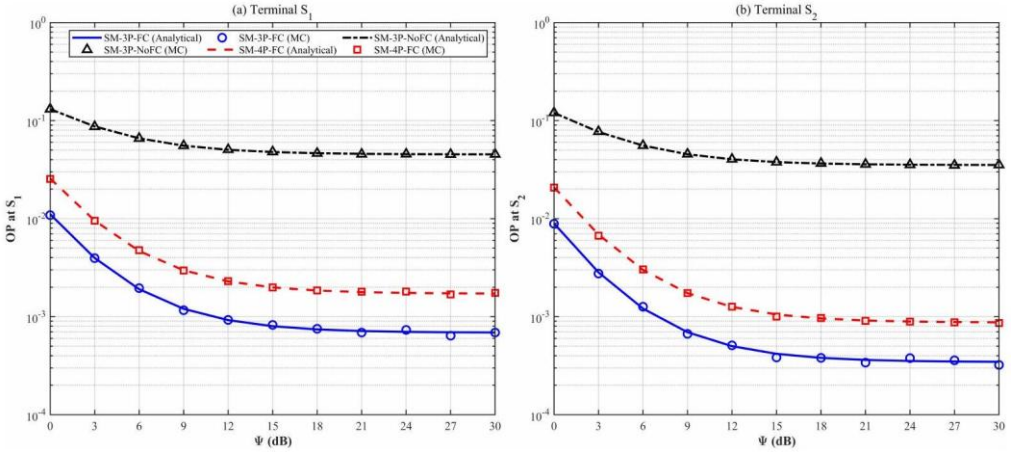


Figure 10. Outage probability of the SM-3P-FC, SM-3P-NoFC and SM-4P-FC schemes *versus* Ψ at (a) S_1 and (b) S_2, with $C_{\text{th}} = 0.1\text{bit/s/Hz}, H = 3, Q = 5, x_R = 0.35$ and $K = 2$.

Over the entire considered normalized transmit-power range, the SM-3P-FC scheme achieves the lowest OP at both terminals, whereas the SM-3P-NoFC benchmark exhibits the highest OP. The ordering is therefore $\text{OP}^{\text{SM-3P-FC}} < \text{OP}^{\text{SM-4P-FC}} < \text{OP}^{\text{SM-3P-NoFC}}$.

The comparison between SM-3P-FC and SM-3P-NoFC directly demonstrates the recovery advantage provided by fountain coding, whereas the comparison between SM-3P-FC and SM-4P-FC quantifies the reliability gain obtained by reducing the number of transmission phases. As Ψ increases, all three schemes approach non-zero outage floors, because the desired and interfering powers are increased proportionally under $P_S = P_I = P$.

Figure 8 has quantified the effect of the number of co-channel interferers. To complement that result and isolate the effect of the interference-power scaling rule, Figure 11 compares the proposed SM-3P-FC scheme under interference-free, fixed-power CCI and proportional-power CCI conditions. The analytical curves closely agree with the MC markers under all three interference conditions. The maximum absolute MC-analytical deviation among the six evaluated cases is below 2.25×10^{-4}, confirming the validity of the derived expressions for unequal values of Ψ_S and Ψ_I, as well as for the interference-free case $K = 0$. At high desired-node powers, several empirical OP values are zero, because no outage event is observed among the $N_{\text{MC}} = 10^6$ channel realizations. These zero-valued MC points are therefore omitted from the logarithmic plots.

Figure 11 demonstrates that both the presence of CCI and the manner in which its power scales with the desired-node transmit power have substantial effects on the outage behavior. In the interference-free case, the OPs at both terminal nodes decrease rapidly as Ψ_S increases, because only thermal noise and the desired-link fading affect packet delivery.

When the interference power is fixed at $\Psi_I = 5$ dB, the OPs are higher than those in the interference-free case, particularly in the low- and moderate- Ψ_S regions. Nevertheless, the OPs continue to decrease as Ψ_S increases. This behavior follows from $\omega_1 = \frac{\rho_{\text{th}}\Psi_I}{\Psi_S} \to 0$ and $\omega_2 = \frac{\rho_{\text{th}}}{\Psi_S} \to 0$, when Ψ_I remains fixed and Ψ_S increases.

By contrast, when the desired and interfering powers increase proportionally; i.e., $\Psi_I = \Psi_S$, the ratio Ψ_I/Ψ_S remains constant and $\omega_1 = \rho_{\text{th}}$. Consequently, the improvement obtained by increasing Ψ_S gradually saturates and the OPs approach non-zero interference-limited floors. Across the entire considered range, the interference-free case provides the lowest OP. When $\Psi_S < 5$ dB, the fixed-power CCI case may exhibit a higher OP than the proportional-power case, because the fixed interference power exceeds the desired-node power. The two CCI conditions would coincide at $\Psi_S = 5$ dB, where the fixed interference power equals the desired-node transmit power. For $\Psi_S > 5$ dB, the fixed-power CCI becomes progressively weaker relative to the desired signal, whereas the proportional-power CCI remains equally strong in relative terms and ultimately produces a non-zero outage floor.

Across both terminal nodes and all three interference scenarios, the global maximum absolute deviation between the analytical and Monte Carlo outage probabilities is approximately $2.243 \times$

10^{-4}. This maximum deviation occurs at terminal S_2 under the fixed-power CCI setting. The close agreement between the analytical and Monte Carlo results supports the accuracy of the derived outage-probability expressions over the evaluated desired-power range.

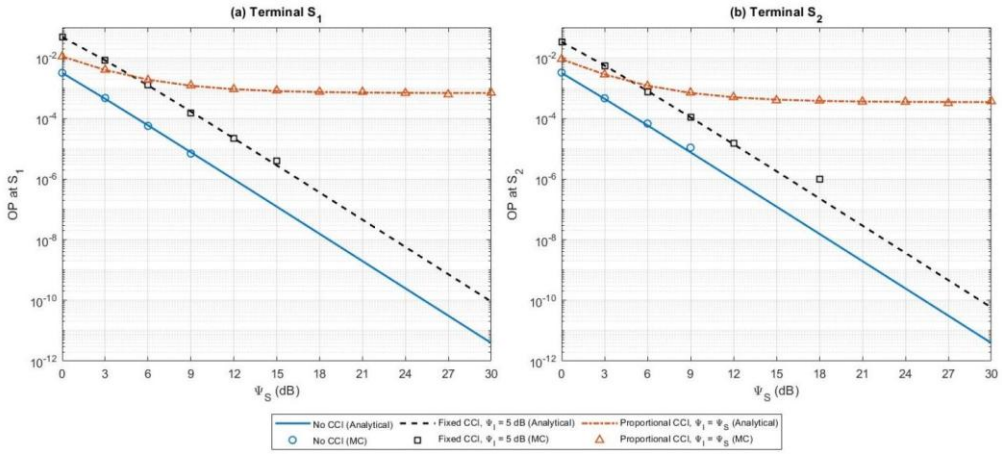


Figure 11. Effect of CCI power scaling on the outage probability of SM-3P-FC at (a) S_1 and (b) S_2: no CCI ($K = 0$), fixed CCI ($K = 2, \Psi_I = 5$ dB) and proportional CCI ($K = 2, \Psi_I = \Psi_S$).

At relatively high desired-node powers, several Monte Carlo outage estimates in the interference-free and fixed-power CCI cases become zero, because no outage event is observed among the $N_{\mathrm{MC}} = 10^6$ generated channel realizations. Since zero-valued samples cannot be displayed on a logarithmic scale, the corresponding MC markers are omitted from Figure 11. These zero-valued estimates reflect the finite Monte Carlo resolution rather than an analytical claim of exactly zero outage probability.

Engineering design implications: The numerical results provide several guidelines for the design of fountain-coded two-way relay networks operating under co-channel interference. First, increasing the desired-node transmit power cannot eliminate the outage floor when the interference power follows the same scaling. Under proportional scaling of P_S and P_I, the ratio Ψ_I/Ψ_S remains constant and the system remains interference-limited. By contrast, when P_S increases while P_I remains bounded, the desired-to-interference power ratio improves and the outage probability continues to decrease. Therefore, desired-node power allocation should be combined with interference-aware scheduling and interference suppression rather than uniformly increasing the powers of all transmitting nodes.

Second, the relay position should be selected by considering both desired link distances and the exposure of the relay and source nodes to the co-channel interferers. Placing the relay closer to one source can improve the corresponding hop while degrading the opposite communication direction, which may produce unequal outage probabilities at S_1 and S_2. Hence, a balanced relay placement should minimize the worst outage probability of the two end nodes rather than optimizing only one link.

Third, the fountain-code parameters H and Q introduce a reliability-latency trade-off. For a fixed H, increasing Q gives the destinations more opportunities to collect the required encoded packets and therefore reduces the outage probability, but it also increases the maximum delivery delay and the number of transmission rounds. In contrast, for a fixed Q, increasing H imposes a stricter recovery requirement and generally increases the outage probability. Thus, H and Q should be jointly selected according to the required reliability and latency constraints of the considered WSN application.

Finally, the performance degradation caused by a larger number of co-channel interferers indicates that the analyzed SM-3P-FC scheme is most beneficial when combined with appropriate interference control. The three-phase structure reduces the packet exchange by one time slot compared with the conventional four-phase protocol, but its practical reliability still depends on the number and locations of the interferers, the relay placement and the selected fountain-code parameters. The reported results are conditioned on the fixed interference geometry specified in Table 3. In deployments with randomly located interferers, the reliability may also vary substantially across spatial realizations, because the relay and the two terminal nodes can experience different aggregate interference levels.

4. CONCLUSIONS AND FUTURE DIRECTIONS

This paper investigated the outage and sum-throughput performance of a three-phase two-way relay network that combines fountain coding and packet-level digital network coding in the presence of co-channel interference. The considered SM-3P-FC protocol reduces the packet exchange from four time slots in the conventional SM-4P-FC protocol to three time slots by combining the relay broadcast operation with digital network coding. Closed-form expressions for the packet-delivery probabilities and the resulting outage probabilities at both terminal nodes were derived over independent block-Rayleigh fading channels. Based on these results, a bandwidth-normalized bidirectional sum-throughput expression was also developed to characterize the reliability-rate trade-off.

The analytical results were validated through Monte Carlo simulations using 10^6 independently generated channel realizations per plotted point. The close agreement between the analytical curves and simulation markers supports the accuracy of the derived outage-probability and throughput expressions. The numerical results show that the SM-3P-FC scheme consistently achieves lower outage probabilities and higher sum throughput than the SM-4P-FC scheme under the considered system configurations. When the desired and interfering transmit powers increase proportionally, the system becomes interference-limited and the outage probabilities approach non-zero floors. The additional CCI-power comparison further shows that the outage probabilities continue to decrease when interference is absent or when its power remains fixed, whereas proportional desired/interference power scaling produces non-zero outage floors. The results further demonstrate that relay placement, the recovery threshold H, the maximum number of exchange rounds Q and the number of co-channel interferers K have significant effects on system reliability. In addition, the sum throughput first increases and then decreases with the target spectral efficiency, revealing an optimum operating rate for each protocol.

These findings provide several practical design implications. Increasing the desired-node transmit power alone cannot eliminate the outage floor when the interference power follows the same scaling. Relay placement should therefore be selected by considering the reliability of both communication directions rather than optimizing only one link. Hence, the coding strategy, number of transmission phases and fountain-code parameters H, Q and C_{th} should be jointly selected according to the required reliability, latency and throughput constraints.

Under the common Figure 9 setting and over the evaluated target-spectral-efficiency grid $C_{\mathrm{th}} \in \{0.1, 0.2, \ldots, 1.0\}$ bit/s/Hz, the maximum evaluated bidirectional sum throughputs of the SM-3P-FC, SM-3P-NoFC and SM-4P-FC schemes are approximately 0.434, 0.289 and 0.322 bit/s/Hz, respectively. Accordingly, within the considered numerical setting, the maximum evaluated throughput of SM-3P-FC is approximately 50.3% higher than that of SM-3P-NoFC and approximately 34.7% higher than that of SM-4P-FC. The outage-probability comparison further shows that SM-3P-FC achieves the lowest outage probability among the three schemes over the considered normalized transmit-power range. These findings separately demonstrate the benefits of fountain coding and the reduction in the number of transmission phases.

Coding-assisted cooperative communication has also been investigated for WSNs using LDPC/OFDM approaches [39], supporting the broader relevance of channel coding in resource-constrained cooperative networks.

The present analysis adopts an ideal packet-level fountain-coding model and assumes fixed node locations and independent block-Rayleigh fading. Equal desired and interfering transmit powers are used in Figures 3-10, whereas Figure 11 considers interference-free, fixed-interference power and proportional-power settings. Future work may consider specific practical fountain-code constructions and degree distributions, finite-length decoding overhead, effective recovery thresholds and decoding-complexity-aware performance analysis for resource-constrained WSN nodes. Other useful extensions include separate source and relay power allocation together with interference-management mechanisms under energy and interference constraints; binomial- and Poisson-point-process models for randomly distributed co-channel interferers, together with spatially averaged outage and throughput analyses; more general fading models, such as Rician and Nakagami- m fading; imperfect channel-state information; spatially correlated channels; and multi-relay extensions employing sum-rate-based relay selection and adaptive relay placement.

Prior studies have shown that sum-rate-based selection can outperform max-min or end-to-end selection in the corresponding half/full-duplex two-way relaying models [40][41][42]. Future extensions should also account for feedback delay and channel-estimation error, since these impairments may reduce the achievable diversity order and coding gain. Security-oriented extensions may further investigate the impact of feedback delay on the secrecy performance of half/full duplex two-way relaying [43]. Additional communication metrics, such as delay, energy efficiency, bit-error rate and ergodic capacity, may also be considered.

REFERENCES

[1] J. N. Laneman, D. Tse and G. W. Wornell, "Cooperative Diversity in Wireless Networks: Efficient Protocols and Outage Behavior," IEEE Transactions on Information Theory, vol. 50, no. 12, pp. 3062-3080, 2004.

[2] R. Ahlswede, N. Cai, S.-Y. Li and R. W. Yeung, "Network Information Flow," IEEE Transactions on Information Theory, vol. 46, no. 4, pp. 1204-1216, DOI: 10.1109/18.850663, 2000.

[3] W. Han, S. Xu, D. Huang and C. Xu, "Analysis and Design of Rateless Two-way Relay Networks Based on a Multiply-and-forward Scheme," Applied Sciences, vol. 10, no. 7, Article no. 2389, 2020.

[4] P. N. Son and T. T. Duy, "A New Approach for Two-way Relaying Networks: Improving Performance by Successive Interference Cancellation, Digital Network Coding and Opportunistic Relay Selection," Wireless Networks, vol. 26, no. 2, pp. 1315-1329, DOI: 10.1007/s11276-019-02186-1, 2020.

[5] D. R. Pattanayak, V. K. Dwivedi and V. Karwal, "Physical Layer Security of a Two Way Relay Based Mixed FSO/RF Network in the Presence of Multiple Eavesdroppers," Optics Communications, vol. 463, Article no. 125429, DOI: 10.1016/j.optcom.2020.125429, 2020.

[6] T.-T. T. Nguyen and D.-T. Do, "Exploiting Performance of Two-way Non-orthogonal Multiple Access Networks: Joint Impact of Co-channel Interference, Full-duplex/Half-duplex Mode and SIC Receiver," Ad Hoc Networks, vol. 97, Article no. 102032, DOI: 10.1016/j.adhoc.2019.102032, 2020.

[7] P. M. Nam et al., "Outage Performance of Interference Cancellation-aided Two-way Relaying Cognitive Network with Primary TAS/SC Communication and Secondary Partial Relay Selection," Electronics, vol. 11, no. 22, Article no. 3645, DOI: 10.3390/electronics11223645, 2022.

[8] V. Ozduran and E. Soleimani-Nasab, "Multi-user Two-way Relaying in the Presence of Co-channel Interference," IET Communications, vol. 14, no. 21, pp. 3853-3865, DOI: 10.1049/iet-com.2020.0170, 2020.

[9] T. T. Duy et al., "Outage Probability of Interference Cancellation Based Two-way Relaying Cognitive Radio Protocol with Primary MIMO Communication," Proc. Of the 2020 IEEE 8th Int. Conf. on Communications and Electronics (ICCE), pp. 481-486, DOI: 10.1109/ICCE48956.2021.9352135, 2021.

[10] D. Sun, Z. Liu and X. Zhang, "Dynamic Cooperative Communications with Mutual Information Accumulation for Mobile Robots in Industrial Internet of Things," Sensors, vol. 24, no. 13, p. 4362, 2024.

[11] N. Q. Sang et al., "On the Security and Reliability Trade-off of the Satellite Terrestrial Networks with Fountain Codes and Friendly Jamming," EAI Endorsed Transactions on Industrial Networks and Intelligent Systems, vol. 10, no. 4, e3, DOI: 10.4108/eetinis.v10i4.4192, 2023.

[12] G. Huang et al., "Improving Throughput in SWIPT-based Wireless Multi-relay Networks with Relay Selection and Rateless Codes," Digital Communications and Networks, vol. 10, no. 4, pp. 1131-1144, 2024.

[13] H. Shao, H. Zhu and J. Bao, "Analysis and Design of Enhanced Distributed Fountain Codes in Multiple Access Networks with Cooperative Relay," Symmetry, vol. 14, Article no. 2026, DOI: 10.3390/sym14102026, 2022.

[14] V. T. Nguyen et al., "On Performance Evaluation of Hybrid Satellite-terrestrial Relaying Networks with Fountain Coding, NOMA and RIS," REV J. on Electronics and Communi., vol. 15, no. 3, pp. 113-124, 2025.

[15] N. V. Toan et al., "Security-reliability Analysis of NOMA-assisted Hybrid Satellite-terrestrial Relay Multi-cast Transmission Networks Using Fountain Codes and Partial Relay Selection with Presence of Multiple Eavesdroppers," EAI Endorsed Trans. On Industrial Networks and Intelligent Systems, vol. 12, no. 3, DOI: 10.4108/eetinis.v12i3.8604, 2025.

[16] S. Atapattu et al., "Reconfigurable Intelligent Surface Assisted Two-way Communications: Performance Analysis and Optimization," IEEE Trans. On Communications, vol. 68, no. 10, pp. 6552-6567, 2020.

[17] J. Wang, Y.-C. Liang, J. Joung, X. Yuan and X. Wang, "Joint Beamforming and Reconfigurable Intelligent Surface Design for Two-way Relay Networks," IEEE Trans. on Communication, vol. 69, no. 8, pp. 5620-5633, 2021.

[18] H. Sun, F. Han, S. Zhao and H. Deng, "Optimal Energy-harvesting Design for AF and DF Two-way Relay Beamforming in 6G," EURASIP J. on Wireless Communi. And Networking, vol. 2022, Article no. 73, 2022.

[19] K. Zhong and L. Fu, "Throughput Maximization for the Full-duplex Two-way Relay System with Energy Harvesting," Electronics, vol. 12, no. 1, Article no. 16, DOI: 10.3390/electronics12010016, 2023.

[20] H. T. Nguyen et al., "Fountain Coding Based Two-way Relaying Cognitive Radio Networks Employing Reconfigurable Intelligent Surface and Energy Harvesting," Telecom, vol. 6, no. 1, Article no. 1, 2025.

[21] C. Bian, Y. Shao, H. Wu and D. Gündüz, "Deep Joint Source-channel Coding over the Relay Channel," Proc. Of the 2024 IEEE Int. Conf. on Machine Learning for Communication and Networking (ICMLCN), pp. 139-144, DOI: 10.1109/ICMLCN59089.2024.10624800, 2024.

[22] C. Bian et al., "Process-and-forward: Deep Joint Source-channel Coding over Cooperative Relay Networks," IEEE Journal on Selected Areas in Communications, vol. 43, no. 4, pp. 1118-1134, 2025.

[23] T. T. Duy and H. Y. Kong, "Exact Outage Probability of Cognitive Two-way Relaying Scheme with Opportunistic Relay Selection under Interference Constraint," IET Comm., vol. 6, no. 16, pp. 2750-2759, 2012.

[24] T.-A. Ngo et al., "Half-duplex Two-way Relaying for Wireless Sensor Networks with Adaptive Coding Rate: A Performance Optimization Framework," Journal of Telecomm. And Inf. Tech., vol. 102, no. 4, pp. 69-76, 2025.

[25] D.-H. Ha et al., "Security and Reliability Analysis of a Two-way Half-duplex Wireless Relaying Network Using Partial Relay Selection and Hybrid TPSR Energy Harvesting at Relay Nodes," IEEE Access, vol. 8, pp. 187165-187181, DOI: 10.1109/ACCESS.2020.3030794, 2020.

[26] N. T. Hau, P. M. Nam, P. T. Tin, T. T. Duy and T. Hanh, "Outage Performance of Two-way Relaying Secure Multicast Networks Using Fountain Codes and Digital Network Coding," Proc. of the 7th Int. Conf. on Green Techn. And Sustainable Development (GTSD), pp. 528-532, DOI: 10.1109/GTSD62346.2024.10674819, 2024.

[27] N. V. Toan, N. T. Hau, P. M. Nam, P. N. Son and T. T. Duy, "Analysis of Security-reliability Tradeoff of Two-way Hybrid Satellite-terrestrial Relay Schemes Using Fountain Codes, Successive Interference Cancelation, Digital Network Coding, Partial Relay Selection and Cooperative Jamming," Telecom, vol. 7, no. 1, Article no. 5, DOI: 10.3390/telecom7010005, 2026

[28] A. Shokrollahi, "Raptor Codes," IEEE Transactions on Information Theory, vol. 52, no. 6, pp. 2551-2567, DOI: 10.1109/TIT.2006.874390, 2006.

[29] M. Luby, "LT Codes," Proc. Of the 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS), pp. 271-280, DOI: 10.1109/SFCS.2002.1181950, 2002.

[30] M. Haenggi, Stochastic Geometry for Wireless Networks, Cambridge University Press, DOI: 10.1017/CBO9781139043816, 2012.

[31] V. Ozduran, J. M. Cioffi and S. B. Yarman, "Opportunistic Source-pair Selection (OSPS) Method for Multiuser Bi-directional Wireless Relaying Networks," Proc. Of the 2013 IEEE 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 565-569, DOI: 10.1109/SPAWC.2013.6612113, 2013.

[32] V. Ozduran, E. Soleimani-Nasab and B. S. Yarman, "Opportunistic Source-pair Selection for Multi-user Two-Way Amplify-and-forward Wireless Relaying Networks," IET Comm., vol. 10, no. 16, pp. 2106-2118, 2016.

[33] V. Ozduran, B. S. B. Yarman and J. M. Cioffi, "Opportunistic Source-pair Selection Method with Imperfect Channel State Information for Multiuser Bi-directional Relaying Networks," IET Communications, vol. 13, no. 7, pp. 905-917, DOI: 10.1049/iet-com.2018.5129, 2019.

[34] V. Ozduran, N. H. Mahmood and N. Nomikos, "Performance Analysis of Power-domain Non-orthogonal Multiple Access for Full-duplex Two-way Relaying," Transactions on Emerging Telecommunications Technologies, vol. 33, no. 7, Article no. e4487, DOI: 10.1002/ett.4487, 2022.

[35] V. Ozduran, "Secrecy Achievable Rate of Two-way Illegitimate Full-duplex Relay-assisted NOMA with CCI," Wireless Personal Communications, vol. 132, pp. 865-888, DOI: 10.1007/s11277-023-10640-9, 2023.

[36] V. Ozduran et al., "Impact of Co-channel Interference on Sum-rate-based Opportunistic Relay Selection (SR-ORS) Method for a Dual-hop Multiple Full-duplex Bi-directional Wireless Relaying Networks," Proc. Of the 2016 National Conf. on Electrical, Electronics and Biomedical Engineering (ELECO), pp. 437-441, 2016.

[37] A. Goldsmith, Wireless Communications, ISBN-13: 978-0-521-83716-3, Cambridge University Press, 2005.

[38] T. S. Rappaport, Wireless Communications: Principles and Practice, 2nd Edn., Prentice Hall, 2002.

[39] V. Ozduran, "LDPCC and OFDM Based Cooperative Communication for Wireless Sensor Networks," Balkan J. of Electrical and Computer Engineering, vol. 8, no. 3, pp. 201-208, DOI: 10.17694/bajece. 688294, 2020.

[40] V. Ozduran, "Sum-rate Based Hybrid Half/Full-duplex Relay Selection in a Dual-hop Two-way Wireless Relaying Networks," Proc. Of the 2017 IEEE Int. Conf. on Sensing, Communication and Networking Workshops (SECON Workshops), pp. 1-5, DOI: 10.1109/SECONW.2017.8314983, 2017.

[41] V. Ozduran, E. Soleimani-Nasab and B. S. Yarman, "Sum-rate Based Relay Selection with Feedback Delay for a Multiple Half/Full-duplex Two-way Relaying," Proc. of the 2017 15th Int. Conf. on ITS Telecommunications (ITST), pp. 1-7, DOI: 10.1109/ITST.2017.7972196, 2017.

[42] V. Ozduran et al., "Sum-rate Based Opportunistic Relay Selection with Channel Estimation Error for a Dual-hop Multiple Half/Full-duplex Bi-directional Wireless Relaying Networks," Proc. of the 2017 26th Wireless and Optical Communication Conf. (WOCC), pp. 1-6, DOI: 10.1109/WOCC.2017.7929009, 2017.

[43] V. Ozduran, B. S. Yarman and J. M. Cioffi, "Feedback Delay Effects on Physical Layer Security in a Dual-hop Half/Full-duplex Bi-directional Wireless Relaying Networks," Proc. of the 2016 IEEE Globecom Workshops (GC Wkshps), pp. 1-7, DOI: 10.1109/GLOCOMW.2016.7848907, 2016.

[44] N. V. Toan, N. N. Lan, T. T. Duy, P. N. Son and N. T. Hieu, "Fountain Codes-based Hybrid Satellite-terrestrial Relay Multicast Schemes in Co-channel Interference Environment: Outage Performance, Joint Time and Power Allocations," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 12, no. 2, pp. 166-182, DOI: 10.5455/jjcit.71-1769912165, 2026.

ملخص البحث:

تبحث هذه الورقة في أداء شبكة ترحيل ثنائي الاتّجاه مخصّصة لشبكات الاستشعار اللّاسلكية، تجمع بين تقنيتي ترميز "النّافورة" وترميز الشبكة الرّقمي في ظلّ وجود تداخل في القناة المشتركة. تمّ النّظر في بروتوكول إرسال مكوّن من ثلاثِ مراحل ومقارنته بالبروتوكول التّقليدي المكوّن من أربعِ مراحل ويستخدم ترميز النّافورة. ولعزل مساهمة ترميز النّافورة، تمّ تقديم نموذجٍ مرجعي إضافي يعتمد على ثلاث مراحل دون ترميز (مع تكرار ثابت) وبالميزانية نفسها، حيث يتمّ توزيع فُرص الإرسال (Q) على (H) من الحزم غير المرمّزة والمتميزة.

وفي ظلّ ظروف تلاشي الإشارة المستقلّ بنمط "بلوك-رايلي"، تمّ اشتقاق تعبيراتٍ تحليلية بصيغةٍ مغلقة لاحتمالات انقطاع الخدمة عند العقدتيْن الطّرفيتيْن، مع أخذ التّداخُل في القناة المشتركة عند كل من عُقدةِ التّرحيل وعُقدة المصدر بعين الاعتبار. وتظهر عمليات محاكاة "مونت كارلو" التي اعتمدت على مليون حالة للقناة تمّ توليدها بشكلٍ مستقلّ لكلّ نقطة بيانية، توافقاً وثيقاً مع التّعبيرات التّحليلية المشتّتة.

وتبيّن النّتائج العدّدية أنّ المخطّط المقترح يحقّق احتمالات انقطاعٍ أقلّ ومجموع إنتاجية ثنائي الاتّجاه أعلى مقارنة بالنّماذج المرجعية. علاوة على ذلك، تُظهر المقارنة أنّ احتمالات الانقطاع تستمر في الانخفاض عند غياب التّداخل أو عند بقاء قدرته ثابتة، في حين أنّ التحجيم النّسبي لقدرة الإشارة المطلوبة مقابل قدرة التّداخل يؤدّي إلى ظهور مستوياتٍ دنيا لاحتمالية الانقطاع بقيمٍ غير صفرية.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

BREAST LESION CLASSIFICATION: LEAKAGE-FREE PATCH-BASED TRANSFER LEARNING WITH CROSS-DATASET EVALUATION ON CBIS-DDSM, MIAS AND INBREAST

Hitesh Kag[1], Dinesh Kumar[2] and Anjali Diwan[3]

(Received: 3-Jul.-2026, Revised: 23-Aug.-2026 and 27-Aug.-2026, Accepted: 28-Aug.-2026)

ABSTRACT

High reported accuracies on small mammographic datasets are frequently inflated by data leakage, most often from augmentation applied before the train/test split. Changing only the split order on whole-image Mini-MIAS moves a deep-feature plus Support Vector Machine (SVM) pipeline from 0.976 accuracy to 0.513, which is equivalent to the random chance. The predictive signal survives at the lesion-patch scale. Our contribution is protocol-level rather than architectural: a reproducible leakage audit; the first quantification and repair of the reported patient overlap in the CBIS-DDSM official split, which we trace to its mass and calcification partitions being drawn independently (31 patients, 8.5% of test lesions); an architecture sweep that tests rather than assumes the performance ceiling; and a label-free procedure for transferring the decision threshold to a new institution. A lightweight EfficientNet patch pipeline trained on CBIS-DDSM under a strict patient-wise protocol obtains ROC-AUC 0.745 (95% CI 0.703–0.794); without retraining, it attains 0.734 (0.640–0.826) on MIAS and 0.877 (0.804–0.946) on INbreast and a DeLong test cannot distinguish the MIAS result from the internal one ($p = 0.82$). Results are stable over ten seeds ($\sigma \leq 0.011$) with clustered bootstrap intervals. Across 11 configurations spanning six architecture families, test AUC spans only 0.707–0.777. For small datasets, leakage control and external validation matter more than headline accuracy.

KEYWORDS

Breast lesion classification, Data leakage, Transfer learning, External validation, Mammography.

1. INTRODUCTION

Breast cancer is the most common cancer and a leading cause of cancer-related death among women worldwide and early detection remains the single most effective means of reducing mortality [1]. X-ray mammography, the principal population-screening modality, is difficult to interpret, because lesions are often small, of low contrast and partly obscured by overlapping fibro-glandular tissue; sensitivity therefore depends heavily on the reader's workload and experience [2]. These challenges have driven decades of computer-aided detection and diagnosis (CAD) research, a fast-moving field that has accelerated further with the advent of deep learning [3].

Convolutional neural networks (CNNs) currently dominate research on mammographic-image analysis. Because well-annotated mammograms are scarce, the prevailing approach is transfer learning, in which a backbone pre-trained on a large corpus of natural images is used as a feature extractor or fine-tuned to the target task. A systematic review of recent medical-imaging studies confirms a strong trend toward feature-extraction and fine-tuning methods, but also notes inconsistent reporting of configurations [4]. In mammography specifically, end-to-end CNNs with tailored pre-processing and augmentation have been shown to classify pre-segmented masses as benign or malignant [5], while lightweight hybrids that combine CNN features with classical classifiers remain attractive for smaller datasets [6].

Despite this progress, reported performance on small public datasets is often difficult to replicate and, on closer inspection, optimistic. Mini-MIAS studies, for example, frequently report high accuracies—such as 93.9% for normal-*versus*-abnormal classification [7] and 96% sensitivity at AUC 0.99 for normal/benign/malignant classification [8]—yet these are typically whole-image results obtained without explicit control of data leakage. One commonly overlooked problem is the application of augmentation before the train/test split: augmented instances (e.g., rotations) of a single image end up

1. H. Kag is with the Department of Computer Eng. (AI, ML & DS), Marwadi Uni., Rajkot, India. Email: hitesh07kag@gmail.com
2. D. Kumar is with the School of Artificial Intell., Bennett Uni. (The Times Group), Greater Noida, India. Email: dineshkr@ieee.org
3. A. Diwan is with the Department of Computer Engineering, Marwadi University, Rajkot, India. Email: anjali.diwan@ieee.org

in both sets, so the images are no longer independent and the reported scores present an over-optimistic picture [9]. Reproducibility and test–retest robustness are likewise rarely assessed, even though small changes in the input can significantly alter CNN predictions [10] and external (cross-dataset) validation—the strongest evidence of generalizability in clinical settings— remains relatively scarce [11]. In this work, we take a different approach: rather than chasing a headline accuracy figure, we focus on methodologically sound evaluation and quantify what is genuinely generalizable. Our contributions are as follows.

- **A reproducible data-leakage demonstration.** We isolate augment-before-split leakage in a controlled experiment that changes one variable. Holding the films, labels, features, classifier and rotation count fixed, splitting the augmented pool at random yields 0.976 accuracy and 0.997 ROC-AUC on whole-image Mini-MIAS, whereas splitting by source film first yields 0.513 accuracy and 0.500 ROC-AUC — chance and below the majority-class rate. We release a single script that runs both arms.
- **A quantified and repaired leakage channel in the CBIS-DDSM official split.** Zafari et al. [12] report that patient overlap exists between the pre-defined CBIS-DDSM training and test sets and discard that split for the purpose of dataset harmonization. We confirm the overlap independently and extend it in four ways: we identify the mechanism; namely that the mass and calcification partitions were drawn independently, so 31 patients contribute a mass case to one side and a calcification case to the other; we quantify it as 60 of the 704 official test lesions (8.5%), with 19 of the 31 patients sharing the same breast and view; we measure its effect on a patch-based transfer-learning pipeline; and we supply a repaired, strictly patient-disjoint split that preserves the official test set, so that results remain comparable with published CBIS-DDSM numbers.
- **A whole-image *versus* lesion-patch analysis.** We show that, under a strict leakage-free protocol, whole-image transfer learning performs far worse than lesion-scale patches for the datasets, backbones and input resolutions evaluated here—now demonstrated on CBIS-DDSM as well as Mini-MIAS—whereas lesion-scale patches carry meaningful and statistically significant signals.
- **Leakage-free cross-dataset external validation.** We train benign-*versus*-malignant classifiers on CBIS-DDSM and test them, without retraining, on MIAS (cross-institution) and INbreast (cross-modality, full-field digital mammography), reporting clustered bootstrap 95% confidence intervals, DeLong tests between datasets and classifier stability across ten random seeds.
- **A tested rather than assumed performance ceiling.** Eleven configurations spanning six architecture families and 2.2–86 M parameters, including ViT-B/16 and Swin-T, are trained under one identical leakage-free protocol. Heavier backbones do help—ViT-B/16 is significantly ahead of the lightweight baseline—but the entire family spans only 0.07 AUC, far short of the accuracies reported under leaky evaluation. The same design also separates backbone from input resolution and, in doing so, corrects a claim in the submitted version.
- **A label-free procedure for transferring the operating threshold.** We find that dis- crimination survives a change of institution, but calibration does not, so a threshold tuned at the source site misbehaves at the target. We give a rank-based transfer that repairs the operating point using only un-labelled target scores and show that on the like-for-like external cohort it recovers the operating point an oracle threshold would choose. We also identify and report, the condition under which it should not be used.

The remainder of the paper is organized as follows. Section 2 reviews related works; Section 3 details the datasets; and Section 4 describes the methodology, including patch extraction, pre-processing, models, the leakage-free protocol, the statistical evaluation and the threshold-transfer procedure. Section 5 presents the results and discusses them. Section 6 details the released reproducibility package. Section 7 tackles the study limitations. Finally Section 8 addresses the conclusions and future work.

2. RELATED WORKS

Deep and transfer learning. Although traditional CAD techniques for mammography were developed over the preceding decade, deep learning has since emerged as the most successful approach, as recent surveys make clear [3]. Under data scarcity, the most common strategy is transfer learning from ImageNet-pre-trained CNNs: such models, combined with augmentation, classify pre-segmented masses as benign or malignant [4], [13] and hybrid approaches classify masses as benign or malignant

or distinguish soft-tissue from bone tumours [5]. On small datasets, however, combining a CNN with a classical classifier (e.g., SVM or dRVFL) still tends to outperform the alternatives [6, 14].

Benchmarks. CBIS-DDSM is an officially train/test-split sub-set of DDSM, curated by using biopsy outcomes to select the best cases of the original DDSM; it was created specifically, because earlier results were difficult to replicate for lack of a standard evaluation set [15]. INbreast provides full-field digital mammography (FFDM) with precise annotations and BI-RADS assessments [16], while larger resources, such as VinDr-Mammo, further extend the landscape [17]. Mini-MIAS, although digitized and relatively small, remains popular and frequently yields strong reported results [7, 8]; its apparent performance, however, depends heavily on the evaluation protocol and its rigour. Attention has also turned to the quality of the benchmarks themselves. Zafari et al. [12] harmonized CBIS-DDSM, TOMPEI-CMMD and VinDr-Mammo into a common format, quantified distributional shifts in breast density and abnormality prevalence between them and reported patient overlap between the pre-defined CBIS-DDSM training and test sets, which led them to discard that split. Their concern is cross-dataset harmonization; ours is what the overlap costs; a classifier that keeps the official split, which we address in Sub-section 5.3.

Patch-based *vs.* whole-image. Whole-image models generally require substantially larger datasets and/or detection-and-aggregation pipelines, whereas patch- or ROI-based classification concentrates the lesion signal and is more effective with limited data [5]. Patch size and resolution matter: a multi-scale design achieved an AUC of ~0.72 on CBIS-DDSM, compared with ≈0.70 for a single-scale design [18].

Augmentation and leakage. For small datasets, augmentation is the most common remedy [19], but when it is applied before splitting, transformed copies of a single image can appear in both partitions, inflating apparent performance. This leakage is rarely controlled for and seldom reported, even when very high accuracies are claimed (e.g., 99.8% on small mammogram sets [20]) [8, 9].

External validation and recent models. Cross-dataset generalization is under-reported, yet crucial: stand-alone models trained to match radiologists in-distribution can perform poorly at different sites or on different scanners, motivating domain-generalization methods [11]. In recent years (2024–2026), research has advanced beyond simple CNNs along three main directions. First, by capturing long-range dependencies, Vision Transformers (ViTs) and hybrid CNN–Transformer models have been applied to tasks, such as BI-RADS and lesion classification, though they typically require thousands of images or extensive pre-training to train reliably—for example, a custom ViT for BI-RADS assessment [21], dual-branch ViT fusion for micro-calcifications [22] and comparative CNN-*versus*-ViT studies for mammography [23]. Second, vision–language foundation models, pre-trained on large collections of mammogram–caption pairs, can improve data efficiency and robustness when data is limited [24], but they perform best on large downstream sets. Third, explainable AI has become an expectation, with methods, such as Grad-CAM [25] and gradient-based attribution [26] routinely used across breast-cancer pipelines and dedicated reviews discussing their use and assessment in mammography [27].

These approaches are generally data-hungry and comparatively heavy to train and deploy. Our focus is complementary: a lightweight, rigorously validated baseline the contribution of which lies in honest, externally validated evaluation rather than architectural novelty. Specifically, our work stands apart in that it is not optimized for peak accuracy, but instead provides (i) a reproducible leakage demonstration, (ii) an explicit whole-image-*versus*-patch comparison and (iii) statistically quantified cross-institution and cross-modality external validation.

3. DATASETS

We use three public mammography datasets (Table 1). CBIS-DDSM [15] is the principal training and testing dataset, providing biopsy-confirmed benign and malignant lesion ROIs with an official train/test partition. The submitted version described that partition as “patient-disjoint”; the word has been struck here, because Sub-section 5.3 shows that the property does not hold once mass and calcification cases are pooled. The leakage study is conducted on Mini-MIAS, which also serves as a cross-institution external test set; lesion centres and radii are read directly from its annotations. INbreast [16] provides full-field digital mammography (FFDM) images and is used as a cross-modality external test set. All labels are mapped to two classes. For CBIS-DDSM, BENIGN and BENIGN_WITHOUT_CALLBACK are mapped to benign and MALIGNANT to malignant, so that the benign-*versus*-malignant separation

follows directly from the biopsy-confirmed labelling. For IN- breast, labels are BI-RADS-derived (1–3 benign, 4–6 malignant) rather than biopsy-confirmed as in CBIS-DDSM—a caveat quantified in Section 5.10.

Table 1 reports the exact composition of every partition at the patient, image, lesion and patch levels. Because each annotated lesion contributes exactly one network input, the lesion and patch counts coincide throughout and only the former is tabulated.

Table 1. Exact composition of every data partition: patients, source images and annotated lesions. Each lesion yields one input patch. The validation fold is carved out of the CBIS-DDSM training split by patient and is re-drawn for every run, so its size is given as a range over the ten seeds.

Dataset	Role	Patients	Images	Lesions	Benign	Malig.
CBIS-DDSM	official train	1248	2458	2863	1682	1181
of which	validation fold	188	—	385–461	—	—
CBIS-DDSM	official test	349	399	704	428	276
Mini-MIAS	external test	104	111	119	68	51
INbreast	external test	50	107	132	43	89

Patient-level independence. Independence was verified programmatically rather than assumed and the audit script is released with the code. Across all ten runs, no patient is ever shared between a training fold and its validation fold. Mini-MIAS films are distributed as left/right pairs, so patient identity is recovered as refnum/2; INbreast redacts patient identifiers in the public release, so its 107 images map onto 50 anonymization tokens, which we treat as patients for clustered resampling. Both external sets are held out in their entirety and never seen in training, so no patient can cross that boundary by construction. The audit did, however, expose one genuine violation inside the CBIS-DDSM official split itself, which we quantify and repair in Sub-section 5.3.

The three datasets are deliberately chosen to probe complementary axes of generalization. CBIS-DDSM, the largest of the three, ships with an official train/test split and biopsy-confirmed labels, eliminating two of the most common confounds—label noise and split arbitrariness—and is therefore adopted as our primary training set. Mini-MIAS, acquired on different equipment, but in the same modality (digitized screen-film), provides a cross-institution stress test, since its site is held out entirely. A more difficult, cross-modality challenge is posed by INbreast: its full-field digital mammograms (FFDMs) have different dynamic ranges, bit depths and noise characteristics from digitized film, so a model that transfers here must cope with a fundamentally different image- formation process. In this setting, we train a single model on CBIS-DDSM and test it—without retraining—on each external set in turn. Statistically, this zero-adaptation protocol disentangles two questions that are conflated in single-dataset studies: robustness to site (the Mini-MIAS axis) and robustness to acquisition technology (the INbreast axis), the latter of which cannot be assessed when training and test data is drawn from the same distribution. Finally, all three datasets are publicly available, so the entire benchmark—the training data, both external test sets and our splits—can be reproduced by any reader, preserving the full reproducibility of the evaluation.

4. METHODOLOGY

The pipeline (Figure 1) is deliberately lightweight, allowing a clean separation of the experimental variables of interest: leakage control, image-*versus*-patch framing, backbone, resolution and cross-dataset transfer.

4.1 Patch Extraction

The network input is a patch rather than a whole mammogram—specifically, a square patch centred on a lesion. For CBIS-DDSM, cropped-ROI images are provided directly. For MIAS, the lesion centre and radius are loaded from the metadata; because MIAS uses a bottom-left origin, the row coordinate is flipped (row = H-y) before cropping, which is necessary to avoid missing the lesion. The crop side length is scaled according to the marked lesion area. For INbreast, mass contours are extracted from the

per-image XML and the resulting bounding box is enlarged by 30% to provide surrounding context. Representative patches are illustrated in Figure 2.

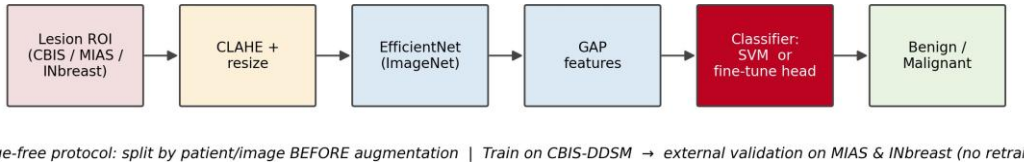


Figure 1. Leakage-free cross-dataset pipeline: lesion ROI→CLAHE→EfficientNet features→classifier, trained on CBIS-DDSM and externally validated on MIAS and INbreast.

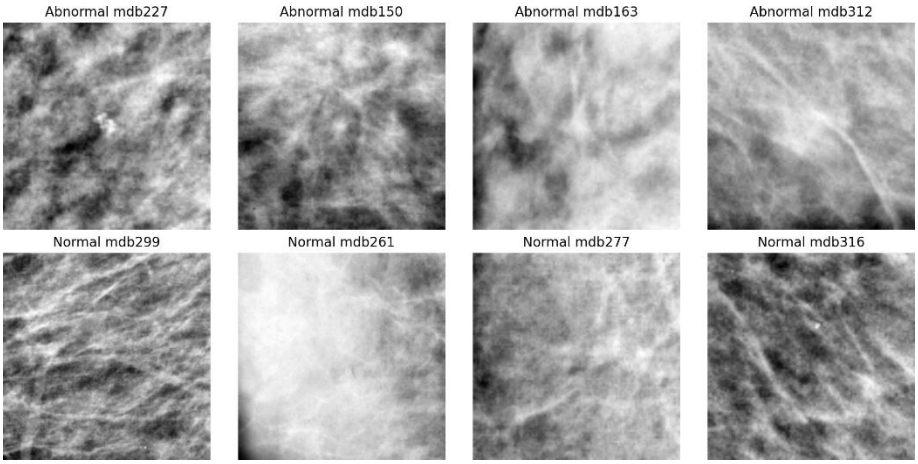


Figure 2. Representative lesion ROI patches after CLAHE enhancement.

4.2 Pre-processing

Each patch is converted to grayscale, resized to the network input size, contrast-enhanced with CLAHE (clip limit 2.0, 8 × 8 tiles) [28] and duplicated across three channels.

Normalization proceeds in three explicit stages, which we spell out here because the term is often used loosely in the literature. (i) Bit-depth windowing. Screen-film scans (CBIS-DDSM, Mini- MIAS) are already 8-bit and are used as-is. INbreast supplies 16-bit FFDM, so intensities are windowed to 8 bits by clipping at the 1st and 99th percentiles of each image and rescaling that interval linearly onto [0, 255]; percentile clipping rather than min–max is used because a handful of saturated detector pixels would otherwise compress the entire breast into a few grey levels. (ii) Local contrast enhancement. CLAHE is then applied to the resized patch. CLAHE is a local, histogram-equalization operator: it equalises the histogram within each 8×8 tile and clips the histogram at 2.0 times the mean bin height before equalizing, which limits noise amplification in the uniform fatty regions, after which tiles are recombined by bilinear interpolation. No global histogram equalization is applied. (iii) Intensity scaling. No z-score or min–max normalization is applied by us. Patches are handed to the network in the raw [0, 255] range, because the Keras EfficientNet models begin with their own Rescaling and Normalization layers that expect un- normalized input and internally apply the ImageNet channel statistics; pre-scaling the input would apply that transform twice. For the PyTorch backbones used in the extended architecture comparison of Sub-section 5.5, which carry no such internal layer, the equivalent transform is applied explicitly: division by 255 followed by subtraction of each model's ImageNet channel mean and division by its channel standard deviation.

4.3 Models and Training

The main pipeline uses three ImageNet-pretrained backbones (MobileNetV2, EfficientNetB0 and EfficientNetB3). These were chosen because compound scaling gives the EfficientNet family a markedly better accuracy-per-parameter trade-off than earlier CNNs at the scale of data available here [29], because the same family has been reported to work well on comparably small medical- imaging sets—including an improved EfficientNetB3 for melanoma lesion classification [30] — and because the whole study had to run on a single 6 GB consumer GPU. We acknowledge, however, that three members

of one efficiency-oriented family cannot by themselves establish that performance is limited by the data rather than by model capacity. Sub-section 5.5 therefore extends the comparison to ResNet-50 [31], DenseNet-121 [32], ConvNeXt-T [33], ViT-B/16 [34] and Swin-T [35] under one identical protocol, so that the claim is tested and not assumed.

We evaluate the three main backbones under two regimes. **Frozen + SVM:** the frozen backbone with global average pooling yields a fixed-length descriptor, which is standardized (zero mean, unit variance) and classified with an RBF-SVM using balanced class weights. **End-to-end fine-tuning:** a head (GAP → dropout 0.4 → sigmoid) is trained in two phases—phase 1 trains the head only (Adam, 10^{-3}, 5 epochs) and phase 2 unfreezes the top ~40–50 layers and continues from a low learning rate (10^{-5}–3×10^{-5}) with early stopping on validation AUC. All batch-normalization layers are kept in inference mode, which stabilizes training with small batches. Training-data augmentation (rotation/flip) is applied on the fly.

4.4 Leakage-free Evaluation Protocol

The key to preventing leakage is that every train/test split is performed before augmentation and at the patient or imaging-unit level, never at the level of individual augmented patches. This guarantees that augmented copies derived from the same source image can never be separated across the split. Figure 3 contrasts the two orderings step by step. The exact splitting rule depends on the dataset. For CBIS-DDSM we split by patient using StratifiedGroupKFold (or GroupShuffleSplit) and use the dataset's official train/test split for the headline results. For Mini-MIAS, we split by image ID, keeping all rotated versions of an image within the same fold. Augmentation is applied only to the training folds, never to the held-out data. The external datasets, Mini-MIAS and INbreast, are never seen during training and are used solely for evaluation.

4.5 Statistical Evaluation

The primary metric is ROC-AUC, which is threshold-independent; accuracy, sensitivity (malignant recall) and specificity are additionally reported at the 0.5 operating point. Where no official split exists, we use five-fold cross-validation (patient-wise for CBIS-DDSM, image-wise for MIAS). We emphasize ROC-AUC as the primary endpoint precisely, because its threshold-independence allows comparison across datasets and operating points. Hyper-parameters are taken from the literature at standard default values; a leakage-safe search (with no use of the test set for selection) confirms that they are near-optimal (Sub-section 5.7).

Three distinct sources of uncertainty are separated, because conflating them is a common cause of overstated claims.

(i) Training variability is measured by repeating the entire two-phase procedure with ten random seeds, which govern both weight initialization of the classifier head and the patient-wise validation carve-out. We report the mean and standard deviation of the per-seed AUCs together with their range. Ten seeds replace the three used in the original submission; three runs bound the spread only very loosely.

(ii) Finite-sample uncertainty of the test set is quantified by percentile bootstrap over 2,000 re-samples. We state the resampling unit explicitly, since it materially changes the width of the interval. Patches drawn from one patient are not independent, so the headline intervals resample *clusters*—patients for CBIS-DDSM, source images for Mini-MIAS and INbreast—taking all patches belonging to each drawn cluster. Patch-level intervals are also reported for comparison and are, as expected, slightly narrower. All intervals are computed on seed-averaged predicted probabilities. We note that a bootstrap can only re-weight the cases actually collected; it cannot represent the between-cohort variability that a genuinely new external sample would exhibit, so for the two small external sets the reported intervals should be read as a lower bound on total uncertainty.

(iii) Differences between AUCs are assessed with DeLong's method [36] and not by inspecting whether an interval excludes chance, using the fast rank-based implementation of Sun and Xu [37]. For two models scored on the same cases-for example two backbones on the CBIS-DDSM test split-we use the paired DeLong test on correlated ROC curves. For two different case sets- for example internal CBIS-DDSM against external Mini-MIAS-the curves are independent and we combine the two DeLong variance estimates additively, $z = (A_1 - A_2)/\sqrt{\mathrm{var}_1 + \mathrm{var}_2}$. Only this comparison can support or refute a claim that external performance matches internal performance; a confidence interval lying

above 0.5 establishes merely that a model beats random guessing, which is a much weaker statement.

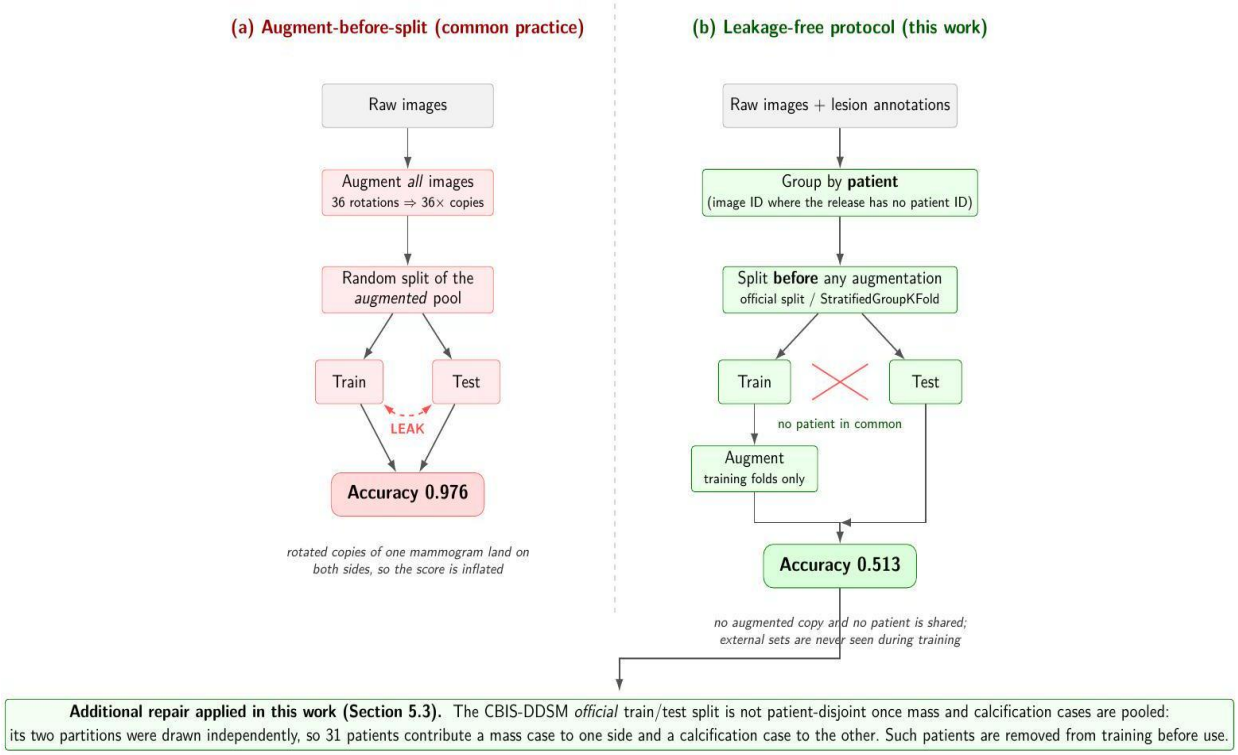


Figure 3. The leakage-free evaluation protocol, with the controlled whole-image Mini-MIAS experiment of Table 16 shown on each branch. (a) When augmentation precedes splitting, rotated copies of the same mammogram appear in both partitions and accuracy reaches 0.976. (b) Grouping by source film and splitting *before* augmenting, then augmenting only the training folds, removes the coupling; the identical pipeline then scores 0.513, which is chance. The lower panel records a further leakage channel, found during this revision, inside the CBIS-DDSM official split itself.

Calibration is reported separately from discrimination, using the Brier score, the expected calibration error over ten equal-width bins and reliability curves, following standard practice for assessing whether modern networks' confidences can be read as probabilities [38]. This distinction turns out to matter: Sub-section 5.9 shows that ranking transfers across domains while calibration does not.

4.6 Label-free Cross-domain Threshold Transfer

Our own results show that a model can rank lesions correctly on a new scanner while its probability scale drifts, so that a threshold tuned on the source institution no longer sits where it should sit. A deployed second reader would encounter exactly this failure mode and reporting AUC does not address it. We add a small procedure that repairs the operating point without requiring labelled data from the target site.

Let S_{src} be the scores the trained model assigns to the held-out *source* validation fold and let $t^{\star}$ be the threshold maximizing Youden's J [39] on that fold. Both are available before deployment. At the target site we observe only un-labelled scores S_{tgt}. We map each target score onto the source score distribution by rank,

$$\tilde{s} = F_{\text{src}}^{-1}\big(F_{\text{tgt}}(s)\big), s \in S_{\text{tgt}}, \tag{1}$$

where F_{tgt} is the empirical CDF of the target scores and F^{-1} the empirical quantile function of the source scores and then apply the unchanged source threshold $t^{\star}$ to $\tilde{s}$. Because the mapping in (1) is monotone, it leaves the ranking and therefore the ROC curve and AUC exactly unchanged; it moves only the operating point. The procedure needs no target labels, no retraining and no access to the source images—only the source score distribution, a few hundred numbers that can be shipped with the model.

Equation (1) implicitly assumes that source and target have comparable class prevalence, since matching quantiles matches the flagged fraction. Where that assumption fails—INbreast is 67% malignant against 41% for CBIS-DDSM—we also evaluate a prevalence-aware variant that rescales the flagged fraction by the ratio of target to source prevalence, which requires an estimate of the target prevalence, but still no case-

level labels. Both are compared in Sub-section 5.9 against a fixed 0.5 threshold and against an oracle that tunes Youden's *J* directly on the labelled target set, the latter serving as an upper bound that a deployed system could not attain.

5. RESULTS AND DISCUSSION

5.1 The Leakage Effect (Mini-MIAS)

The submitted version of this table compared an augment-before-split accuracy against leakage- free ROC-AUCs that had been produced by different scripts on a different task and asserted that "only the augmentation/split order differs". That was not accurate: between those rows, the label definition, the sample, the backbone and the rotation count all changed as well. We have replaced it with a properly controlled experiment.

Table 16 holds everything constant—the same 115 Mini-MIAS films, the same benign-*versus*-malignant labels, the same frozen MobileNetV2 features, the same RBF-SVM, the same 36 rotations — and changes exactly one thing: whether the split is taken before or after augmentation. Splitting the augmented pool at random gives accuracy 0.976 and ROC-AUC 0.997. Splitting by source film first and augmenting only the training side, with the identical pipeline, gives accuracy 0.513 and ROC-AUC 0.500—below the 0.556 majority-class rate and indistinguishable from chance. The entire apparent skill of the first configuration is an artefact of rotated copies of one film appearing on both sides of the split.

The mechanism is straightforward. With 36 rotations per film and a random split of the augmented pool, almost every test sample has a near-duplicate of itself in the training partition, so even a nearest-neighbor lookup would score highly. The effect is a specific instance of a phenomenon documented across machine-learning-based science [9], in which pre-processing or augmentation applied before partitioning couples the partitions and yields over-optimistic, irreproducible results. The safeguard is simple and frequently neglected: partition at the imaging-unit level, by image or preferably by patient, before any augmentation, is applied.

Under the leakage-free protocol, the whole-image task remains non-discriminative across configurations. Table 2 reports three further leakage-free variants on the related normal-*versus*-abnormal task — frozen features, end-to-end fine-tuning and 512 px inputs—none of which recovers any skill. We label these explicitly as a different task from Table 16, rather than tabulating them alongside the benign-*versus*-malignant numbers as though they were comparable.

Table 2. Leakage-free whole-image Mini-MIAS, *normal versus abnormal* (322 films). These represent a different task from Table 16 and are not comparable with it as numbers; they are reported to show that the absence of whole-image skill under a leakage-free protocol is not an artefact of one backbone or one input resolution. Majority-class accuracy for this task is 0.64.

Configuration	Metric	Score	Note
Frozen, 224 px	ROC-AUC	0.518 ± 0.083	chance
Fine-tuned, 224 px	ROC-AUC	0.471	chance
Frozen, 512 px	ROC-AUC	0.478 ± 0.023	chance

5.2 Lesion-patch Signal Is Real (Mini-MIAS)

Re-framing the problem at the lesion scale (image-wise, leakage-free, normal *versus* abnormal) recovers a well-separated discrimination (Figure 4A). The suspiciously high AUC of the naïve normal-patch sampling arises, because the normal patches are not necessarily lesion-sized or lesion-centred, whereas the abnormal patches are. When normal samples are resampled to match the size distribution of the abnormal patches and given a tissue-density-weighted location, the AUC drops by about 0.10 to 0.841 ± 0.026, which we term the *honest* AUC. The fact that a large margin above chance persists even under this stringent control indicates that the residual signal is genuinely pathological rather than an artefact of patch selection. The per-fold breakdown (Table 3) shows that discrimination is consistent across the five image-disjoint folds (AUC 0.806–0.868), with sensitivity for malignant/abnormal lesions (~0.67) being the limiting factor—as expected, given that only 119 abnormal lesions were available. This pattern, in which sensitivity rather than specificity is the bottleneck, recurs throughout our experiments and is discussed further below. We stress that this normal-*versus*-abnormal AUC (0.841) is a *different* task from the benign-*versus*- malignant

external validation reported in Sub-section 5.8 (MIAS AUC 0.734); the two numbers use different label sets and are not directly comparable.

Table 3. Mini-MIAS lesion-patch (size/density-matched), per-fold image-wise 5-fold CV.

Fold	Acc	AUC	Sens.	Spec.
1	0.803	0.868	0.680	0.878
2	0.846	0.864	0.739	0.905
3	0.779	0.828	0.615	0.881
4	0.828	0.837	0.565	0.976
5	0.794	0.806	0.727	0.829
Mean	**0.810**	**0.841**	**0.665**	**0.894**

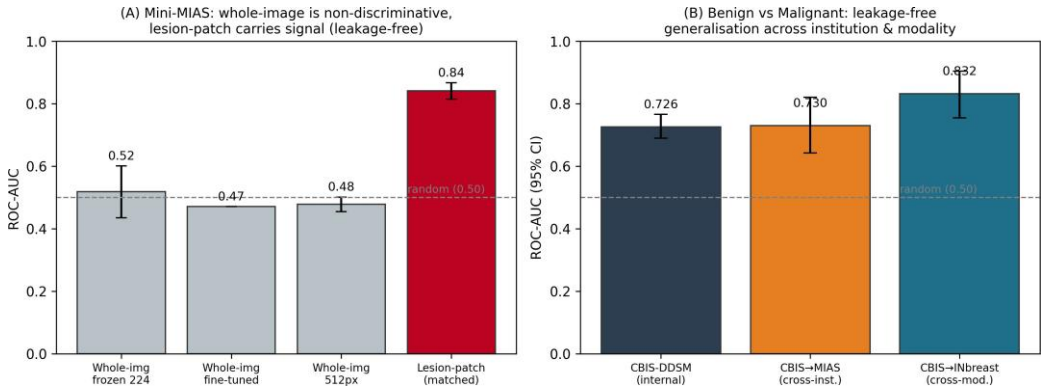


Figure 4. (A) On Mini-MIAS, whole-image classification is non-discriminative under a leakage-free protocol, whereas lesion patches carry signal. (B) Benign-vs-malignant: leakage-free generalization across institution (MIAS) and modality (INbreast), with 95% confidence intervals.

5.3 The CBIS-DDSM Official Split Is Not Patient-disjoint

Applying the patient-independence audit requested during review to our own protocol identified a violation, described below.

We are not the first to observe that this overlap exists. Zafari et al. [12], in the course of harmonizing several mammography datasets, noted patient overlap between the pre-defined CBIS-DDSM training and test sets and discarded the split for that reason. Our contribution here is not the observation, but what follows from it: the mechanism that produces the overlap, its size, its measured effect on a patch-based transfer-learning pipeline and a repaired split that keeps the official test set usable rather than abandoning it.

The submitted manuscript described the CBIS-DDSM official train/test partition as patient- disjoint. For the way the present study and much of the literature, uses it, that description does not hold. The mass and calcification sub-sets were partitioned independently of one another. Each is internally clean: we found zero patients shared between train and test within the mass sub-set and zero within the calcification sub-set. However, 31 patients have both a mass lesion and a calcification lesion and for those patients, the two draws disagreed, placing one lesion type in training and the other in test. Pooling mass and calcification cases into a single benign-*versus*-malignant task, as we and many others do, reintroduces the patient-level dependence that the official split appears to guarantee.

The scale is modest, but not negligible. Sixty of the 704 official test lesions (8.5%) belong to these 31 patients. In 19 of the 31 cases, the shared patient contributes the same breast and the same view to both sides, so the model may encounter the same breast tissue, with the same parenchymal pattern and acquisition characteristics, in training and at test time. The lesions themselves are distinct, as are the image files: we verified that no cropped ROI, no full mammogram and no pixel array is duplicated across the split. The dependence lies at the patient and tissue levels and is not a duplicated-sample error.

We quantify the consequence by retraining the identical model on a repaired split. The repair keeps the official test set untouched, so results remain comparable with published CBIS-DDSM numbers, instead removing every affected patient from training: 83 training lesions are dropped, leaving 2,780 and the residual train/test

patient overlap is zero. Table 10 reports both arms over ten seeds.

The measured effect is negligible. Internal AUC moves by +0.001 (from 0.745 to 0.746) and a paired DeLong test on the shared 704 test lesions does not distinguish the two models ($p = 0.68$). The external results are likewise unchanged and the finding that external performance matches internal performance holds in both arms. We state the null result plainly, since a leak of this kind might otherwise be assumed to have inflated the reported figures; on this dataset, with 8.5% of test lesions affected and the shared material being one breast and not one lesion, it has not.

Three observations nonetheless follow. First, the contamination is a property of a widely used public benchmark and not of our pipeline and its magnitude in another study, for example one pooling differently, using a smaller test set or fitting a higher-capacity model, is not bounded by the value measured here. Second, any study pooling mass and calcification cases from CBIS- DDSM should apply this repair or split by patient directly, since the official split alone does not provide the guarantee its documentation implies. Third, the contamination entered through a dataset convention, not a modelling error and is detectable only by generating and auditing split manifests. This supports treating split manifests and not only source code, as a standard released artefact. That it has been noted independently by Zafari et al. [12] while harmonizing the same benchmark reinforces rather than weakens the point: a property of a widely used dataset went unremarked in the classification literature that relies on it and was surfaced twice, in parallel, only by groups that inspected the split directly.

5.4 CBIS-DDSM Benign *vs.* Malignant: Ablation

For the larger, biopsy-proven CBIS-DDSM, we ablate the pipeline one component at a time on the official test split (Table 4). The only meaningful gains come from fine-tuning (+0.026) and a stronger backbone (+0.027); increasing the input resolution yields no improvement (-0.011), while test-time augmentation recovers part of the loss (+0.013). A leakage-safe hyper-parameter analysis (Sub-section 5.7) confirms that these defaults are near-optimal and the per-dataset confusion matrices are shown in Figure 5. The submitted version continued here with a claim that performance saturates at 0.72–0.73 and that this indicates a task/data ceiling. That sentence has been removed: as the caveats below and the factorial sweep of Sub-section 5.5 show, the single-run ablation could not support it.

Two caveats on how this table should be read. First, each row is a single training run, whereas the headline results of Sub-section 5.8 average ten. Single runs are noisy at this scale: the EfficientNetB0 @224 entry here (0.724) sits at the bottom of the ten-seed range subsequently measured for the same configuration (0.745, range 0.723–0.753), so the two are consistent, but no ordering among rows separated by less than about 0.02 should be trusted. Second, backbone and input resolution change together between the last three rows, so this table cannot attribute the difference to either. Both problems are addressed in Sub-section 5.5, which repeats the comparison with a proper factorial design over multiple seeds and it is that table, not this one, on which we base any claim about a performance ceiling.

Table 4. CBIS-DDSM component ablation (official test split). Single training run per row; see the caveats in the text and the multi-seed factorial comparison in Table 13.

Backbone	Training	Input/TTA Test	AUC
MobileNetV2	Frozen + SVM	224	0.671
MobileNetV2	Fine-tuned	224	0.697
EfficientNetB0	Fine-tuned	224	0.724
EfficientNetB3	Fine-tuned	300	0.713
EfficientNetB3	Fine-tuned	300 + TTA	**0.726**

5.5 Does a Heavier Backbone or a Larger Input Help?

The ablation above varies backbone and input resolution together, leaving the two confounded and all three of its backbones belong to one efficiency-oriented family. Neither is a sound basis for claiming a task/data ceiling, so the experiment was repeated with a design that separates them.

Eleven configurations were trained under the identical leakage-free protocol — same patient-wise splits, same two-phase schedule, same augmentation, same early stopping—spanning classical CNNs (ResNet-

50, DenseNet-121), a modernized CNN (ConvNeXt-T), transformers (ViT-B/16, Swin-T) and the EfficientNet family, from 2.2-86 M parameters. To keep the comparison internally consistent across families, this study runs in a single PyTorch implementation instead of mixing frameworks; EfficientNet-B0 @224 is included in it as a cross-framework anchor against the TensorFlow result of Table 7. EfficientNet-B0, EfficientNet-B3 and ResNet-50 are each run at 224 and 300 px, which gives the clean resolution ablation with the backbone held fixed. Table 13 reports the outcome and Table 11 the paired DeLong tests against the EfficientNet-B0 reference on the shared test cases.

Two results follow, the first of which corrects the submitted manuscript. With the backbone held fixed, resolution does improve performance: EfficientNet-B0 gains +0.036 AUC from 224 to 300 px ($p = 0.003$, paired DeLong) and EfficientNet-B3 gains +0.022. The original ablation reported that resolution brought no benefit (-0.011), but it varied backbone and resolution together; once the two factors are separated, the sign of the effect reverses. The factorial design requested during review is what made this visible.

Second, the performance ceiling is present, but not flat and the claim has been weakened to match. Across 11 configurations spanning 2.2–86 M parameters, CBIS-DDSM test AUC ranges from 0.707 to 0.777. The strongest configuration is ViT-B/16 @224, the advantage of which over the lightweight EfficientNet-B0 baseline is statistically significant (+0.060, $p = 0.0002$); ConvNeXt-T and the higher-resolution EfficientNets are also significantly ahead, while ResNet-50, DenseNet-121, Swin-T and MobileNetV2 are not distinguishable from it (Table 11). Capacity thus retains some influence, in that an 86 M-parameter transformer extracts more than a 4 M-parameter CNN. What the sweep establishes is the scale of the remaining headroom: the full set of architectures, classical and modern, convolutional and attention-based, spans 0.07 AUC and none approaches the 0.94–0.99 figures that leakage-free evaluation dissolves. The difference between a well-evaluated and a leaky protocol is an order of magnitude larger than the difference between the best and the worst architectures.

The ordering also fails to transfer. Swin-T is the strongest model on Mini-MIAS (0.775) while ranking mid-table internally and EfficientNet-B3 is among the weakest externally despite being competitive internally. Selecting an architecture on internal performance alone would not have identified the configuration that generalizes best, which constitutes a further argument for routine external validation.

The anchor proved necessary. An initial PyTorch implementation reproduced it approximately 0.12 AUC too low, with external transfer falling to near chance. The cause was batch normalization, not the architecture: the TensorFlow pipeline holds every BN layer in inference mode, whereas PyTorch by default continues to update BN running statistics during training, which on batches of 16 to 32 grayscale patches is sufficient to destabilize a 10^{-5} fine-tuning stage. In the absence of the anchor, the resulting table would have measured BN conventions while being presented as a comparison of backbones. We record this, because it illustrates the same point the rest of the paper makes: on small datasets, an implementation detail can outweigh the model choice under study.

5.6 Whole Mammogram *versus* Lesion Patch on the Primary Dataset

The submitted manuscript concluded that lesion patches beat whole-image analysis, but demonstrated it only on Mini-MIAS, the whole-image task of which collapses to chance once leakage is removed. That is a weak basis for a general claim, so we added the missing baseline on CBIS-DDSM itself.
Each full mammogram is labelled malignant if any lesion it contains is malignant, giving 2,857 images (2,458 train / 399 test) under the same patient-wise protocol. We train at 224 px and again at 512 px, because a lesion occupying a few hundred pixels in a 3000×5000 mammogram may simply be destroyed by aggressive down-sampling; if the patch advantage were only a resolution artefact, the 512 px model should close much of the gap. Table 14 reports the comparison against the patch pipeline trained and tested on the same patients.

The patch framing performs better, but the resolution effect is substantial. At 224 px the whole- image model reaches AUC 0.638±0.024, more than 0.10 below the patch pipeline; at 512 px it recovers to 0.705±0.014, closing approximately two-thirds of that gap while remaining about 0.04 short. Part of what the submitted manuscript attributed to framing is thus attributable to resolution: reducing a 3000×5000 mammogram to 224 px removes much of the lesion, which is a more accurate account than the claim that whole images carry no signal. The residual gap at 512 px represents the framing effect properly. It is consistent in direction with the Mini-MIAS result while being considerably smaller, as would be expected

on a dataset an order of magnitude larger.

Two differences make this a comparison of framings and not a strictly controlled experiment. The whole-image task carries an image-level label (malignant if any lesion is malignant) over 399 test images, against 704 lesion-level labels for the patch task and a whole-image model must localize as well as classify. We draw a limited conclusion: at the data scale available here, the patch framing extracts more signal. Nothing in these results implies that whole-image mammography classification is un-learnable.

5.7 Hyper-parameter Sensitivity (Leakage-safe)

Tuning is itself prone to optimistic bias, particularly on small datasets, so our main configurations use the literature-standard defaults, with only the epoch count set by early stopping. To verify that these defaults are not materially sub-optimal, we ran a search that never used the official test set for selection. For the frozen-feature SVM (MobileNetV2 features, as in Table 4), a grid search over C and γ was performed with patient-wise five-fold cross-validation (the test set excluded throughout); the best configuration ($C = 1$, $\gamma = 10^{-4}$) attained a CV-AUC of 0.721—0.008 above the default—and a test AUC of 0.679 (Table 5). For fine-tuning, an unfreeze-depth ×learning-rate search selected a 3×10^{-5} configuration, the test AUC of which was 0.733 being only +0.009 over the default (Table 6), with the test set used for reference only. Tuning produced no meaningful test-set gains for either classifier and the ∼0.72–0.73 ceiling is unchanged, confirming that the reported results do not depend on a lucky choice of defaults.

5.8 Cross-dataset External Validation

The transferred model is **EfficientNetB0 @224**, fine-tuned on the CBIS-DDSM official training split and applied to both external sets without any retraining. We state this explicitly, because the submitted version described it as "the best configuration of Table 4", which was an error: that row is EfficientNetB3 @300 with TTA, a different model that happened to reach an internal AUC agreeing with EfficientNetB0's to three decimal places. EfficientNetB0 is also the more appropriate choice for a transfer experiment, since test-time augmentation would confound domain transfer with an inference-time ensembling effect.

Table 5. Leakage-safe SVM grid (patient-wise CV-AUC on the training split; top rows).

C	γ	CV-AUC
1	10^{-4}	**0.721**
1	scale	0.712
1	10^{-3}	0.708
10	10^{-4}	0.708
0.1	10^{-4}	0.706

Table 6. Leakage-safe fine-tuning sweep (validation-selected; test for reference only).

Unfrozen	Phase-2 LR	Val. AUC	Test AUC
30	10^{-5}	0.772	0.710
30	3×10^{-5}	0.779	0.722
50	10^{-5}	0.779	0.723
50	3×10^{-5}	**0.781**	**0.733**

Table 7 reports the results over ten seeds and Table 11 the significance tests. Three points stand out.

5.9 Calibration and Repairing the Operating Point

Table 9 and Figure 6 separate the two things an AUC hides. Discrimination is preserved on transfer, but the expected calibration error roughly doubles on Mini-MIAS and the reliability curves depart from the diagonal in *opposite* directions on the two external sets. On Mini-MIAS, the curve lies below the diagonal: the model is *over*-confident, assigning malignancy probabilities higher than the observed malignant fraction, which is what drives the specificity loss at a fixed 0.5 threshold. On INbreast, it lies above the diagonal—the model is *under*-confident—which is the signature of a cohort the prevalence of which (67% malignant) is far higher than the source's (39%) rather than of a model that has degraded. That the two external sets fail in opposite directions is itself informative and it is why a single blanket recalibration rule

is not appropriate. A two-sample Kolmogorov–Smirnov test against the source validation scores confirms that the score *distribution* shifts in both cases; usefully for deployment, this statistic needs no target labels, so a site can measure its own drift before deciding whether to trust an inherited threshold.

Table 7. Leakage-free external validation. A single EfficientNetB0 @224 trained on the CBIS-DDSM; official training split is applied without retraining to both external sets. AUC is given as the mean±standard deviation over ten seeds, alongside the observed range; the 95% CI is a cluster bootstrap (2,000 resamples of patients for CBIS-DDSM, of source images for the external sets) on the seed-averaged predictions. Sensitivity and specificity are at the 0.5 operating point.

Test set	n	Clusters	AUC (10 seeds)	Range	95% CI	Sens.	Spec.
CBIS-DDSM (internal test)	704	349	0.745±0.008	0.723–0.753	0.703–0.794	0.70	0.65
CBIS→MIAS	119	111	0.734±0.011	0.713–0.750	0.640–0.826	0.78	0.51
CBIS→INbreast	132	107	0.877±0.009	0.858–0.891	0.804–0.946	0.89	0.74

Table 8. Confusion matrices computed directly from the saved per-sample predictions (seed-averaged, threshold 0.5), replacing the matrices reconstructed from summary statistics in the submitted version. FN is the clinically costly error.

Test set	TP	FN	TN	FP	Sens.	Spec.
CBIS-DDSM (internal test)	194	82	277	151	0.70	0.65
CBIS→MIAS	40	11	35	33	0.78	0.51
CBIS→INbreast	79	10	32	11	0.89	0.74

Table 9. Discrimination transfers, but calibration does not. All quantities are computed on the seed-averaged predicted probabilities, so the AUC column is the seed-ensembled value and differs marginally from the seed-mean of Table 7. AUC is essentially unchanged from internal to external data, whereas the Brier score and expected calibration error degrade and the score distribution shifts away from the source (two-sample Kolmogorov–Smirnov statistics against the source validation scores, computable without any target labels).

Test set	AUC (ens.)	Brier	ECE	KS *vs.* source
CBIS-DDSM (internal test)	0.749	0.206	0.104	0.084
CBIS→MIAS	0.737	0.218	0.153	0.210
CBIS→INbreast	0.882	0.131	0.085	0.293

First, discrimination transfers. The external AUC on Mini-MIAS is statistically indistinguishable from the internal CBIS-DDSM AUC, the DeLong test on independent samples giving $p = 0.82$, so equality cannot be rejected. The submitted manuscript instead argued that the external confidence intervals lay above 0.5 and so demonstrated "statistically significant generalization", which conflated beating chance with matching internal performance. That statement has been withdrawn. The INbreast AUC is significantly *higher* than internal ($p < 0.001$), a property of its labelling and not of the model (Sub-section 5.10).

Table 10. Effect of repairing the CBIS-DDSM official split. Pooling mass and calcification cases makes the official partition non-patient-disjoint: 31 patients appear on both sides, affecting 60 of the 704 test lesions. The repair removes those patients from training (83 lesions) and leaves the official test set untouched, so the two columns are evaluated on identical cases. AUC is the mean over ten seeds.

Test set	Official split	Patient-disjoint	Δ
CBIS-DDSM (internal test)	0.745 ± 0.008	0.746 ± 0.004	+0.001
CBIS→MIAS	0.734 ± 0.011	0.731 ± 0.009	-0.003
CBIS→INbreast	0.877 ± 0.009	0.871 ± 0.008	-0.006

Second, the result is stable across training runs: the standard deviation over ten seeds is ≤ 0.011 on every

test set, with the full range spanning less than 0.03 AUC. We report ten seeds in place of three, because three runs estimate this spread poorly.

Third—and this is the finding with the clearest deployment consequence — discrimination transfers, but calibration does not. At a fixed 0.5 threshold, the ranking is intact while specificity on Mini-MIAS degrades, because the whole score distribution drifts upward relative to the source (Sub-section 5.9). A model shipped with its source threshold would therefore over-flag at a new site while its AUC looked healthy.

Table 11. DeLong tests. Internal and external AUCs are computed on different cases, so their ROC curves are independent and the two DeLong variances are combined additively; comparisons on shared cases use the paired test. The internal-*versus*-MIAS comparison is the one that supports the generalization claim: the difference is not significant; i.e., external performance is statistically indistinguishable from internal performance, which is a stronger statement than a confidence interval merely excluding chance.

Comparison	Type	AUC A	AUC B	z	p
CBIS internal *vs.* MIAS	unpaired	0.749	0.737	0.23	0.818
CBIS internal *vs.* INbreast	unpaired	0.749	0.882	-3.67	<0.001
MIAS *vs.* INbreast	unpaired	0.737	0.882	-2.60	0.009
Official *vs.* patient-disjoint (CBIS internal)	paired	0.749	0.750	-0.41	0.681
Official *vs.* patient-disjoint (MIAS)	paired	0.737	0.733	0.74	0.462
Official *vs.* patient-disjoint (INbreast)	paired	0.882	0.878	0.64	0.522

Table 12 then asks what a deployer should actually do about it, comparing four strategies against an oracle that tunes Youden's J on the labelled target set and is therefore unattainable in practice.

Table 12. Choosing the operating point at a new institution. The oracle tunes Youden's J on the labelled target set and is therefore unattainable in deployment; it is shown as an upper bound. Carrying the source threshold across unchanged is the naive baseline. Quantile transfer and its prevalence-matched variant use only un-labelled target scores (plus, for the latter, an estimate of target prevalence). Best attainable J per dataset in bold.

Threshold strategy	Thr.	Sens.	Spec.	Youden J	Needs labels?
CBIS-DDSM (internal test)					
fixed 0.5 (as submitted)	0.500	0.703	0.647	0.350	no
source Youden J, carried across	0.504	0.692	0.652	0.344	no
quantile transfer	0.504	0.670	0.680	0.350	no
prevalence-matched transfer	0.533	0.659	0.699	**0.358**	no
oracle target Youden J	0.563	0.623	0.762	0.385	yes
CBIS→MIAS					
fixed 0.5 (as submitted)	0.500	0.784	0.515	0.299	no
source Youden J, carried across	0.504	0.765	0.529	0.294	no
quantile transfer	0.504	0.686	0.706	**0.392**	no
prevalence-matched transfer	0.546	0.686	0.706	**0.392**	no
oracle target Youden J	0.546	0.686	0.706	0.392	yes
CBIS→INbreast					
fixed 0.5 (as submitted)	0.500	0.888	0.744	**0.632**	no
source Youden J, carried across	0.504	0.888	0.744	**0.632**	no
quantile transfer	0.504	0.640	0.907	0.547	no
prevalence-matched transfer	0.608	0.775	0.837	0.612	no
oracle target Youden J	0.532	0.888	0.791	0.678	yes

Carrying the source-validated Youden threshold across unchanged, which is what the submitted manuscript

implicitly recommended, performs no better on Mini-MIAS than taking no action at all: Youden's J is 0.294 against 0.299 for a fixed 0.5 threshold. Quantile transfer (Equation 1) raises J to 0.392, equal to the value attained by the oracle (0.392) and the prevalence-matched variant yields the same operating point. Specificity on Mini-MIAS rises from 0.51 to 0.71 for a modest loss of sensitivity, so the specificity collapse that motivated the analysis is largely eliminated without annotating any case at the target site. Mini-MIAS is the like-for-like external cohort and this is the principal result of the section.

One negative result is reported alongside it. On INbreast, the plain 0.5 threshold is already close to the oracle and quantile transfer degrades it. The reason is visible in Table 9: the INbreast score distribution is shifted principally, because its prevalence differs (67% malignant against 39% in the source) rather than because the model is miscalibrated on that cohort. Rank matching forces the target to flag the same fraction of cases as the source, which is the wrong correction under prevalence shift; the prevalence-matched variant recovers most, though not all, of the loss. The resulting practical rule is narrow. The operating point should be transferred when the score distribution has drifted and prevalence is comparable and the prevalence-matched form should be preferred wherever a prevalence estimate is available. We make no claim that the procedure is safe to apply without checking these conditions.

Table 13. Architecture and resolution under one identical leakage-free protocol, run in a single PyTorch implementation so the families are directly comparable. EfficientNet-B0/B3 and ResNet-50 appear at 224 and 300 px, so resolution is varied with the backbone held fixed. AUC is the mean over three seeds on the CBIS-DDSM official test split; external AUCs are for the same models applied without retraining.

Backbone	Params (M)	Input	CBIS test AUC	MIAS	INbreast
ViT-B/16	85.8	224	0.777 ± 0.012	0.736	0.838
ConvNeXt-T	27.8	224	0.749 ± 0.005	0.719	0.839
EfficientNet-B3	10.7	300	0.747 ± 0.008	0.697	0.727
Swin-T	27.5	224	0.742 ± 0.006	0.775	0.837
EfficientNet-B0	4.0	300	0.742 ± 0.009	0.684	0.811
ResNet-50	23.5	224	0.738 ± 0.000	0.720	0.836
ResNet-50	23.5	300	0.731 ± 0.007	0.706	0.836
DenseNet-121	7.0	224	0.725 ± 0.005	0.726	0.841
EfficientNet-B3	10.7	224	0.725 ± 0.007	0.689	0.760
MobileNetV2	2.2	224	0.722 ± 0.020	0.688	0.806
EfficientNet-B0	4.0	224	0.707 ± 0.001	0.738	0.818

Table 14. Whole-mammogram *versus* lesion-patch classification on CBIS-DDSM, the primary dataset, under the same patient-wise leakage-free protocol. Two input resolutions are used, so that the gap cannot be dismissed as an artefact of down-sampling a full mammogram until the lesion disappears. Mean over three seeds.

Framing	Input	n (train/test)	CBIS test AUC
Whole mammogram	224 px	2458 / 399	0.638 ± 0.024
Whole mammogram	512 px	2458 / 399	0.705 ± 0.014
Lesion patch	224 px	2863 / 704	0.745

Table 15. Sensitivity to imperfect lesion localization. Each crop is regenerated from the source image with a random translation, expressed as a fraction of the patch side or with the box scaled; the model is not retrained. Because the crops are regenerated rather than taken from the official pre-cropped ROIs, the top row is the correct baseline for reading the degradation. Mean over three random jitter draws and a three-seed model ensemble.

Translation	Scale	CBIS test	MIAS	INbreast
0%	1.00	0.756	0.734	0.887
5%	1.00	0.755	0.757	0.875
10%	1.00	0.735	0.758	0.878
15%	1.00	0.732	0.715	0.864
20%	1.00	0.722	0.719	0.845
0%	0.85	0.750	0.751	0.875
0%	1.15	0.744	0.760	0.861

Table 16. The leakage effect, measured with a single variable changed. Both arms use the same 115 Mini-MIAS films, the same benign-*versus*-malignant labels, the same frozen MobileNetV2 features, the same RBF-SVM and the same 36 rotations; they differ only in whether the split is taken before or after augmentation. Mean ± standard deviation over 5-fold cross-validation. Majority-class accuracy for this task is 0.556.

Protocol	Split unit	Accuracy	ROC-AUC
Augment → split	augmented sample	0.976 ± 0.006	0.997
Split → augment train only	source film	0.513 ± 0.051	0.500 ± 0.072

Table 17. Quantitative Grad-CAM localization against the radiologist annotation, with two null models scored on the identical patches. Crops are deliberately off-centred and the mask is cropped through the same transform. Pointing is the fraction of patches the peak activation of which falls inside the annotation energy; the fraction of activation mass inside it; conc. the energy divided by the lesion's area fraction, which equals 1 for a uniform map and is therefore comparable across regimes. In the native crop convention the lesion fills too much of the patch for the test to discriminate and the centre prior wins; the wide-field crops, where the lesion covers about a tenth of the patch, represent the informative comparison.

Setting	Saliency	Pointing	Energy	Conc.	IoU@top20	Lesion area
CBIS-DDSM, native crop	Grad-CAM	0.375	0.276	1.27	0.226	0.218
	centre prior	0.876	0.448	2.05	-	
	random	0.084	0.218	1.00	-	
Mini-MIAS, native crop	Grad-CAM	0.639	0.549	1.20	0.214	0.459
	centre prior	0.899	0.686	1.50	-	
	random	0.252	0.459	1.00	-	
CBIS-DDSM, wide field	Grad-CAM	0.272	0.107	1.72	0.123	0.062
	centre prior	0.029	0.105	1.69	-	
	random	0.024	0.062	1.00	-	
Mini-MIAS, wide field	Grad-CAM	0.513	0.295	1.89	0.294	0.156
	centre prior	0.202	0.253	1.62	-	
	random	0.042	0.156	1.00	-	

5.10 Why the INbreast Result Must Be Read Cautiously?

The INbreast AUC is the highest of the three, but this warrants caution: its BI-RADS 1–2 cases are unambiguously normal/benign, making the separation easier than in the biopsy-confirmed CBIS-DDSM. The fair, like-for-like external comparison is therefore CBIS→MIAS; the higher INbreast figure should not be read as one modality being inherently more accurate.

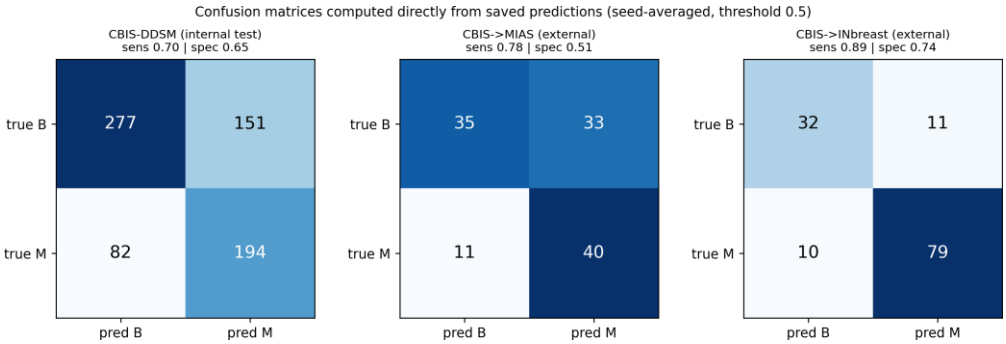


Figure 5. Confusion matrices at the 0.5 operating point, computed directly from the saved per-sample predictions. In the submitted version, these counts were reconstructed from the reported sensitivity and specificity; they are now generated from the prediction files released with the paper. Note the specificity drop on MIAS, caused by un-calibrated cross-domain thresholds rather than by any loss of ranking ability.

Reviewers rightly pressed us to quantify and not merely acknowledge this, so we report the exact BI-RADS composition. Of the 132 INbreast mass patches, 89 fall in the malignant class (BI-RADS 4a–6: 7, 2, 12, 56 and 12 patches for 4a, 4b, 4c, 5 and 6, respectively) and 43 in the benign class. That benign class is *not* homogeneous: 30 patches are BI-RADS 2, genuinely benign, but 13—**30.2% of the benign class**—are

BI-RADS 3, "probably benign". BI-RADS 3 denotes a lesion with a ≲2% malignancy risk that warrants short-interval follow-up rather than biopsy; it is an assessment of *management*, not a histological finding and a minority of such lesions do prove malignant.

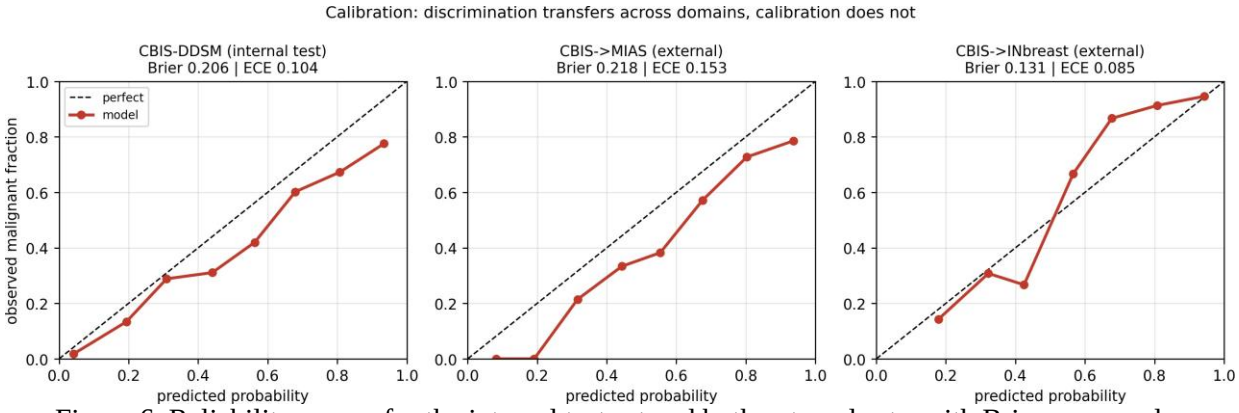


Figure 6. Reliability curves for the internal test set and both external sets, with Brier score and expected calibration error. Points below the diagonal indicate over-confidence. Discrimination is largely preserved across the domain change (Table 7) while calibration is not, which is why a threshold tuned at the source institution misbehaves at the target.

This biases the comparison in two ways, both of which inflate INbreast. First, assignment to BI-RADS 2 *versus* 4–6 is driven by the same morphological features the classifier reads; namely margin, shape and speculation, so the label is partly a function of image appearance rather than an independent ground truth. The task is consequently closer to predicting a radiologist's assessment than to predicting histology, whereas CBIS-DDSM labels derive from biopsy and are independent of appearance. Second, the genuinely ambiguous cases that a biopsy-confirmed cohort contains, in particular benign lesions with malignant appearance, are largely absent, since such lesions would have been assessed BI-RADS 4 and hence assigned to the malignant class. The 13 BI-RADS 3 patches are the only substantial source of label noise acting in the opposite direction. We therefore treat the INbreast figure as evidence of robustness to a change of acquisition technology rather than as a measure of diagnostic accuracy and base all generalization claims on CBIS→MIAS.

A more principled treatment would not force an ordinal scale into two classes at all. BI-RADS is ordered and the distance from 2 to 3 is not the distance from 3 to 4; ordinal regression or soft labels that encode the published malignancy probability of each category (roughly < 2% for BI-RADS 3 against > 95% for BI-RADS 5), would let a model exploit that structure and would make the uncertainty explicit rather than hiding it in a thresholding decision. Vision–language foundation models pre-trained on report text are well placed to do this, since the reports they are trained on contain the assessment category and the hedging language that accompanies it. We leave this to future work, but note that it is the natural way to use BI-RADS-labelled cohorts as external validation without importing their labelling convention as if it were ground truth.

5.11 How Much Localization Error Can the Classifier Absorb?

Every result so far assumes the lesion's centre or bounding box is known exactly, because it is read from the dataset's annotations. A deployed pipeline would receive that box from a detector, with error. To estimate what this costs, we regenerate each crop from the source image with a random translation—5, 10, 15 or 20% of the patch side, in a random direction—and separately with the box scaled by 0.85 or 1.15 to mimic a detector that crops too tightly or too loosely and re-evaluate the trained model without any retraining. Table 15 reports the results on all three datasets.

Because these crops are regenerated rather than taken from the official pre-cropped ROIs, the zero-shift row rather than the headline AUC is the correct baseline for reading the degradation; we would otherwise be attributing the difference between two crop conventions to localization noise.

Degradation is gradual rather than abrupt. A displacement of 15% of the patch side, which corresponds to a realistic detector error, costs 0.024 AUC on CBIS-DDSM, 0.019 on Mini-MIAS and 0.022 on INbreast, each comparable to the seed-to-seed spread of the model itself. Displacements up to 5% have negligible

effect and at 20%, the loss remains within approximately 0.03. Scaling the box by ±15% has less effect than translating it. For an integrator, this indicates that the classifier does not require pixel-accurate boxes and that a detector with typical localization error would not be the binding constraint. The experiment measures only a displaced box on a true lesion; a full detection-plus-classification evaluation would additionally have to account for regions proposed on normal tissue and would yield a lower end-to-end figure.

5.12 Interpretability

We compute Grad-CAM [26] attention maps from the fine-tuned EfficientNet on lesion patches (Figure 7). The submitted version used these panels to *confirm* that predictions are driven by the lesion; that was more than the evidence supported, so we have both softened the claim and measured it.

Measuring it requires care. If each patch is cropped centred on its lesion, then "the peak activation lies inside the annotation" and is satisfied by any model with a centre bias and the metric says nothing. We therefore off-centre every crop and crop the annotation mask through the identical transform, so the lesion sits at a known but non-central position. Against that, we score three quantities—the pointing-game hit rate, the fraction of Grad-CAM energy inside the annotation and the IoU of the top-20% activation with it—and, on the very same patches, two null models: a fixed Gaussian centre prior and a smoothed random map. We also report a chance-normalized concentration, the energy inside the mask divided by the mask's area fraction, which equals 1 for a uniform map regardless of lesion size and so is comparable across regimes.

This precaution proves insufficient on its own. In the crop convention on which the model is trained, the lesion occupies between a fifth and a half of the patch, so it cannot be displaced far enough to leave the central region. The centre prior consequently outscores Grad-CAM, by 0.88 against 0.37 on CBIS-DDSM and 0.90 against 0.64 on Mini-MIAS (Table 17, "native" rows). In this regime, the pointing game provides evidence about the cropping convention rather than about the model, an observation that applies to a substantial body of published saliency evaluation on pre-cropped lesion patches. We therefore repeated the analysis on wide-field crops of three times the lesion box, in which the lesion covers approximately a tenth of the patch and can be placed genuinely off-centre. The ordering then reverses: the Grad-CAM peak falls inside the annotation for 0.27 of CBIS-DDSM patches, against 0.03 for the centre prior and 0.02 for a random map and for 0.51 of Mini-MIAS patches against 0.20 and 0.04. These correspond to approximately nine and two-and-a-half times the centre prior respectively and an order of magnitude above chance.

We draw a bounded conclusion. The maps are consistent with lesion-driven prediction and localize the annotated abnormality considerably better than chance, but on their own, they do not establish that the model reasons from lesion morphology and saliency scores computed on tightly cropped lesion patches should not be read as though they did.

5.13 Comparison with Prior Work

Direct numerical comparison is confounded by differing tasks, splits and (often unreported) leakage control (Table 18). A clear trend nonetheless emerges: the largest headline figures come from the settings most susceptible to optimistic bias. Both report very high single-dataset accuracy without an explicit augment-before-split control—99.8% by Adedigba et al. [20] on a small mammogram set using a discriminatively fine-tuned DenseNet and 98.87% by Salama and Aly [40] using a segmentation-plus-CNN pipeline on DDSM. Methods closer to our own—feature-fusion or deep-feature-plus-classifier approaches—report similarly high single-dataset accuracy, but are likewise never externally validated: Jafari and Karami [41] reach 94.5% on MIAS with an ensemble of CNNs, Ragab et al. [6] obtain 87.2% (AUC 0.94) on DDSM/CBIS-DDSM using AlexNet features with an SVM and Yu et al. [14] fuse deep features with an SVM for an AUC of 0.93 on DDSM. By contrast, our leakage-free CBIS-DDSM AUC (0.745) is close to the only directly comparable patch-based single-dataset CBIS result (0.722) [18] and is unique among these works in providing patient-wise splitting, two-dataset external validation, clustered bootstrap confidence intervals and released per-sample predictions. We therefore present our numbers not as state-of-the-art accuracy, but as a reliable benchmark that most prior studies—lacking leakage control and external validation—cannot offer.

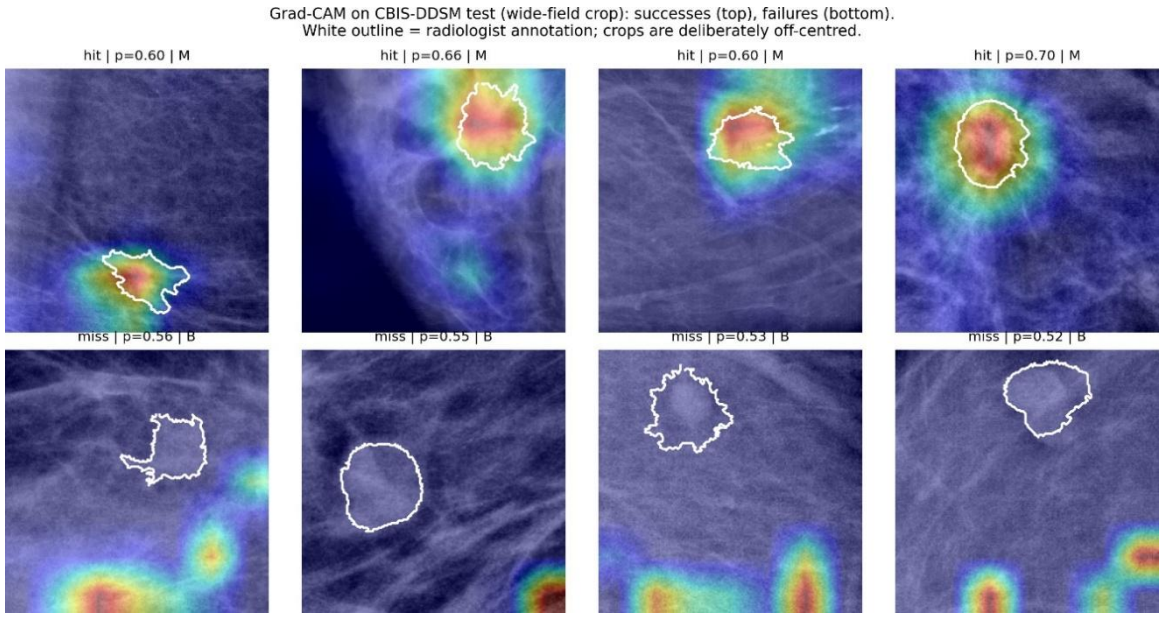


Figure 7. Grad-CAM on deliberately off-centred wide-field CBIS-DDSM crops. Top row: successes. Bottom row: failures, selected automatically as the cases with the least activation mass inside the annotation. The radiologist annotation is outlined in white on every panel, so misses are visible rather than asserted. Every failure shown is a benign lesion scored just above the 0.5 threshold on which the attention has drifted to surrounding fibro-glandular tissue—the false-positive mode discussed in Sub-section 5.15.

Table 18. Contextual comparison with prior work on mammographic lesion classification. Many studies report higher headline numbers, but on whole images or a single dataset and without explicit leakage control or external validation—the dimensions on which this work focuses. The last column states the key limitation relative to our protocol.

Study	Dataset(s)	Method	Reported	Limitation *vs.* this work
Setiawan et al. [7]	MIAS	Law's texture + ANN	93.9% (N/A); 83.3% (B/M)	Whole-image; no leakage control or external validation reported
Agnes et al. [8]	mini-MIAS	Multi-scale CNN (MACNN)	96% sens; AUC 0.99	Whole-image; augment-before-split not controlled; single dataset
Adedigba et al. [20]	Small mammogram set	DenseNet + discriminative fine-tuning	99.8% acc	Near-perfect on small data; standard split, no leakage control or external test
Ragab et al. [6]	DDSM / CBIS-DDSM	AlexNet features + SVM	87.2% acc; AUC 0.94	Image-level; single corpus; no cross-dataset validation
Jafari & Karami [41]	MIAS / DDSM / RSNA	Multi-CNN feature fusion + classifiers	94.5% (MIAS)	Image-level; no patient-wise split; no external validation
Salama & Aly [40]	DDSM / MIAS / CBIS	U-Net segmentation + CNN	98.87% (DDSM)	Segmentation-dependent; very high acc. without reported leakage control
Yu et al. [14]	DDSM	Deep features + dRVFL	AUC 0.93; 81.7% acc	Single dataset; image-level 5-fold; no cross-dataset external validation
Quintana et al. [18]	CBIS-DDSM	Multi-scale (patch)	AUC ≈ 0.722	Patch-based, but single dataset; no external validation
This work	**CBIS (+MIAS, INbreast)**	**EfficientNet patch + SVM/fine-tune**	**AUC 0.745; ext. 0.734 / 0.877**	**Leakage-controlled, patient-wise; 2dataset external validation with clustered 95% CIs and DeLong tests**

5.14 Synthesis

Taken together, the experiments support a single claim: on small mammography datasets, reported performance is governed by evaluation methodology at least as much as by model capacity. The backbone-plus-classifier that appears near-perfect on whole-image Mini-MIAS performs no better than chance once augment-before-split leakage is removed and neither fine-tuning nor higher resolution recovers the apparent skill. The signal that does survive is local: it is recoverable at the lesion-patch scale and withstands a size- and density-matched sampling control. The resulting level is stable across ten seeds,

across a hyper-parameter search, across a repair of the benchmark's own split and across a change of institution. It is not entirely insensitive to capacity, since over 11 configurations spanning 2.2–86 M parameters the test AUC ranges from 0.707 to 0.777, with the largest transformer significantly ahead of the smallest CNN. That range of 0.07 AUC is nevertheless small beside the 0.2 or more separating any of these models from the accuracies that leaky evaluation produces on the same data. The evaluation protocol is thus the larger term. The evidence comes from one lesion-level task on three public cohorts and should not be read as a general law**.**

5.15 Clinical Relevance: What the Errors Would Cost

The intended role of this system is a *second reader* that flags suspicious lesions for radiologist review, never an autonomous diagnostic device. What matters clinically is therefore not the headline AUC, but the asymmetric cost of the two error types and what would have to be true before such a tool could be deployed.

False negatives are the dominant clinical risk. A missed malignancy is the error that changes outcomes: it delays treatment until the next screening round or until the patient presents symptomatically, by which time stage and mortality risk have risen. The per-dataset confusion matrices (Table 8) quantify this: at a 0.5 threshold the model misses 82 of 276 malignant lesions on the internal test set (30%) and 11 of 51 on Mini-MIAS (22%). No operating point available to us makes the system safe to run without a radiologist and a leakage-free figure is what makes this visible; a leaked accuracy of 0.94 would have implied the opposite. In deployment the threshold should therefore be set well below 0.5 to increase sensitivity at the cost of recall workload and the system should be positioned so that its negatives do not suppress a human read.

False positives set the workload cost. Each false positive propagates into recall imaging, patient anxiety and, at the end of the chain, avoidable biopsy of benign tissue. The cross-domain calibration failure matters most here: at a fixed 0.5 threshold, specificity on Mini-MIAS collapses even though the ranking is intact, so a model shipped with its source threshold would flood a new site with false alarms and be switched off long before its discrimination was ever assessed. The transfer procedure of Sub-section 4.6 exists for exactly this reason.

Deployment challenges. Four remain and none is solved here. *(i) Localization.* The pipeline classifies a lesion patch and presumes something has already found the lesion. In a real workflow a detector supplies that box with error; Sub-section 5.11 measures how quickly performance decays as the box degrades and it is the first thing an integrator should re-measure with its own detector. *(ii) Threshold governance.* Because calibration does not transfer, every site needs a defined recalibration step, plus monitoring for drift as scanners and protocols change. The burden here is regulatory and procedural, not only a technical one. *(iii) Case mix.* Our evaluation is lesion-level and enriched with abnormalities; a screening population is overwhelmingly normal, so the positive predictive value in service will be far below anything reported here. *(iv) Scope of evidence.* Both external cohorts are small and retrospective, drawn from curated public archives rather than consecutive clinical series and neither support a claim about a specific device or population.

On the other hand, the computational barrier is genuinely low: the whole pipeline trains and runs on a single 6 GB consumer GPU, so on premise deployment needs no specialized infrastructure and retraining as local annotations accrue is cheap. The binding constraints are clinical and procedural, not computational.

6. REPRODUCIBILITY

Every number in this paper is produced by a script that is released with the manuscript, together with the per-sample predictions those scripts emit. Publishing the raw prediction files matters as much as publishing the code: it is what allows a reader to re-compute our confusion matrices, confidence intervals and significance tests without a GPU and it is what we ourselves lacked when the first version of this manuscript reconstructed its confusion matrices from summary statistics.

The package is available at https://github.com/hitesh712/leakage-free-mammography, archived at DOI:10.5281/zenodo.22122946. Two commands reproduce the statistical results of this paper on a laptop, with no GPU and without downloading the imaging datasets: rev08_analysis.py regenerates the

headline table, the DeLong tests, the calibration figures and the threshold-transfer comparison from the released predictions and rev11_validate_stats.py re-runs the validation of our DeLong implementation reported below.

Released artefacts. *(i)* The full experiment suite: data preparation and patch extraction, the split audit, the ten-seed training run, the patient-disjoint repair, the architecture and resolution ablation, the whole-image baseline, the crop-perturbation study, the Grad-CAM localization analysis and the statistical analysis. *(ii)* The exact split manifests, listing every lesion with its patient identifier, its official partition and its partition after the patient-disjoint repair. *(iii)* Per-sample predicted probabilities for every dataset, every seed and every architecture. *(iv)* The environment records emitted by each run.

Validation of the statistical code. Because the significance claims rest on our own implementation of DeLong's method, we validated it rather than assume it and release the validation script and its output. On the real prediction files, its AUC matches scikit-learn to ten decimal places and its standard error agrees with a 20,000-resample bootstrap standard error to within 0.1%, 1.0% and 0.4% on the internal, Mini-MIAS and INbreast sets, respectively—two independent estimators of the same quantity. Comparing a score vector with itself returns $z = 0$ exactly. Under simulation with two scorers of genuinely equal AUC, the paired test rejects at 0.053 and the unpaired test at 0.046 against a nominal 0.05, so both are correctly calibrated; the unpaired variant matters most here, since it is the less standard of the two and carries the internal-*versus*-external comparison. Power against a large true difference is 0.998 and heavily tied scores are handled correctly.

Settings. Table 19 lists the complete configuration. The ten seeds are 2021, 7, 123, 42, 2024, 5, 777, 31337, 99 and 1234; each seed sets NumPy and framework random state and re-draws the patient-wise validation carve-out. Two software stacks are used and both are reported, since the extended architecture comparison of Sub-section 5.5 runs in PyTorch while the main pipeline runs in TensorFlow.

7. LIMITATIONS

Several caveats apply to these findings. First, the practical ceiling on biopsy-confirmed benign-*versus*-malignant data is malignant sensitivity (Table 8), which would likely require larger cohorts, multi-view fusion or detection-and-aggregation pipelines to improve. Second, INbreast labels are BI-RADS-derived rather than strictly biopsy-confirmed, so its stronger external result reflects greater robustness rather than higher accuracy—the like-for-like comparison remains CBIS→MIAS (Sub-section 5.10). Third, discrimination transfers, but calibration does not, so per- dataset recalibration is required for deployment. Fourth, the approach is intentionally patch-level, relying on per-patch ROI annotations; it does not exploit whole-image triage, multi-view or temporal information and uses a lightweight backbone trainable on a single consumer GPU.

Table 19. Complete training configuration and software environment.

Item	Value
Input	224×224 lesion patch, grayscale replicated to 3 channels, raw [0,255]
Pre-processing	percentile windowing (16-bit only) → resize → CLAHE (clip 2.0, 8 × 8)
Head	global average pooling → dropout 0.4 → dense(1, sigmoid)
Phase 1	head only; Adam, lr10^{-3}; 5 epochs; batch 16
Phase 2	top 40 backbone layers unfrozen; Adam, lr10^{-5}; ≤ 20 epochs
Early stopping	monitor validation AUC, patience 5, restore best weights
Class imbalance	balanced class weights $n/(2n_c)$
Augmentation	rotation ∈ {0, 90, 180, 270} and horizontal flip; training folds only
Validation	GroupShuffleSplit, 15%, grouped by patient, re-drawn per seed
Batch Normalization	kept in inference mode throughout
Seeds	2021, 7, 123, 42, 2024, 5, 777, 31337, 99, 1234
Bootstrap	2,000 resamples, clusters = patients (CBIS) or source images (external)
Main pipeline	Python 3.10.11, TensorFlow 2.10.0, tensorflow-directml-plugin 0.4.0
Architecture study	PyTorch 2.5.1+cu124, timm 1.0.27, mixed precision
Shared libraries	NumPy 1.23.5, pandas 2.3.3, scikit-learn 1.7.2, SciPy 1.15.3, OpenCV 4.13.0, pydicom 3.0.2
Hardware	single NVIDIA RTX 4050 Laptop GPU, 6 GB

Fifth, and most important, is how the external numbers should be read, the external cohorts are small ($n \approx$ 120–130) and our uncertainty estimates understate the true uncertainty. A bootstrap resamples the cases we happen to have collected; it can only express how the estimate would move under re-weighting of *those* cases. It cannot express between-cohort variability—how the estimate would move on a genuinely new sample of patients from a different site, scanner fleet and referral pattern—which for cohorts of this size is likely the larger component. Our intervals are therefore a lower bound on total uncertainty, compounded by both external sets being curated public archives rather than consecutive clinical series. Larger, prospectively assembled cohorts, such as VinDr-Mammo [17] and ideally several of them, are needed before the transfer result can be treated as an estimate of field performance rather than as evidence that transfer occurs at all.

Sixth, the pipeline consumes a lesion patch and therefore presumes that localization has already happened. Sub-section 5.11 measures sensitivity to imperfect boxes, but a full detection-plus-classification evaluation, in which the detector also generates false regions on normal tissue, is outside our scope and would give a lower and more realistic end-to-end figure.

Finally, our claims about whole-image classification are bounded by what we tested: two datasets, EfficientNet-family backbones and input resolutions up to 512 px. We do not claim that whole- image mammography classification is unlearnable in general—with far larger cohorts, multi-view input or detection-and-aggregation architectures demonstrably learnable—only that it did not learn under the data scale and configurations examined here.

8. Conclusion and Future Work

On small mammography datasets, it is methodology rather than architecture that determines reported performance. We identified a widely reproducible leakage pitfall that inflates Mini-MIAS whole-image classification from chance to ~98% and demonstrated leakage-free generalization of a CBIS-DDSM-trained benign-*versus*-malignant classifier (AUC 0.745 internal; 0.734 on MIAS and 0.877 on INbreast, with clustered confidence intervals and ten-seed stability). A DeLong test cannot separate the internal and cross-institution results, which is the specific sense in which the model generalizes.

Two findings sharpen the methodological message. Auditing our own protocol showed that the CBIS-DDSM *official* split is not patient-disjoint once mass and calcification cases are pooled, so leakage entered through a dataset convention rather than a modelling error—an illustration of why split manifests, not just code, should be published. And extending the comparison to six architecture families, including ViT-B/16 and Swin-T, left the performance band unchanged, which converts the ceiling from a conjecture into a tested claim. Finally, because discrimination transfers while calibration does not, we give a label-free procedure that recovers the operating point at a new site.

The result is not a claim of state-of-the-art accuracy, but a reproducible, honestly measured baseline. In future work, we will validate on larger external cohorts, such as VinDr-Mammo [17], investigate domain-generalization and vision–language foundation models—including ordinal or soft-label treatments of BI-RADS—and evaluate the pipeline end-to-end behind a real detector rather than behind ground-truth boxes.

References

[1] A. Bhushan, A. Gonsalves and J. U. Menon, "Current State of Breast Cancer Diagnosis, Treatment and Theranostics," Pharmaceutics, vol. 13, no. 5, p. 723, 2021.

[2] American Cancer Society, "Limitations of Mammograms," [Online], Available: https://www.cancer.org/cancer/types/breast-cancer/screening-tests-and-early-detection/mammograms/limitations-of-mammograms.html, Jun. 2023.

[3] L. Luo et al., "Deep Learning in Breast Cancer Imaging: A Decade of Progress and Future Directions," IEEE Reviews in Biomedical Engineering, 2024, DOI: 10.1109/RBME.2024.3357567.

[4] H. E. Kim et al., "Transfer Learning for Medical Image Classification: A Literature Review," BMC Medical Imaging, vol. 22, p. 69, DOI: 10.1186/s12880-022-00793-7, 2022.

[5] D. Lévy and A. Jain, "Breast Mass Classification from Mammograms Using Deep Convolutional Neural Networks," arXiv Preprint, 1612.00542, [Online], Available: http://arxiv.org/abs/1612.00542, 2016.

[6] D. A. Ragab, M. Sharkas, S. Marshall and J. Ren, "Breast Cancer Detection Using Deep Convolutional Neural Networks and Support Vector Machines," PeerJ, vol. 7, p. e6201, DOI: 10.7717/peerj.6201, 2019.

[7] A. S. Setiawan et al., "Mammogram Classification Using Law's Texture Energy Measure and Neural

Networks," Procedia Computer Science, vol. 59, pp. 92–97, DOI: 10.1016/j.procs.2015.07.341, 2015.

[8] S. A. Agnes et al., "Classification of Mammogram Images Using Multiscale All Convolutional Neural Network (MA-CNN)," J. of Medical Systems, vol. 44, p. 30, DOI: 10.1007/s10916-019-1494-z, 2020.

[9] S. Kapoor and A. Narayanan, "Leakage and the Reproducibility Crisis in Machine-learning-based Science," Patterns, vol. 4, no. 9, p. 100804, DOI: 10.1016/j.patter.2023.100804, 2023.

[10] A. Lemay et al., "Improving the Repeatability of Deep Learning Models with Monte Carlo Dropout," NPJ Digital Medicine, vol. 5, p. 174, DOI: 10.1038/s41746-022-00709-3, 2022.

[11] Y. Yang et al., "MammoDG: Generalizable Deep Learning Breaks the Limits of Cross-domain Multi-center Breast Cancer Screening," arXiv preprint, arXiv: 2308.01057, 2023.

[12] Y. Zafari et al., "MammoClean: Toward Reproducible and Bias-aware AI in Mammography through Dataset Harmonization," arXiv preprint, arXiv Preprint, arXiv: 2511.02400, Nov. 2025.

[13] S. Benbakreti et al., "Using ResNet-18 in a Deep-learning Framework and Assessing the Effects of Adaptive Learning rates in the Identification of Malignant Masses in Mammograms," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 10, no. 1, pp. 93–107, DOI: 10.5455/jjcit.71-1699818406, 2024.

[14] X. Yu, Z. Ren, D. S. Guttery and Y.-D. Zhang, "DF-dRVFL: A Novel Deep Feature Based Classifier for Breast Mass Classification," Multimedia Tools and Applications, vol. 83, no. 5, pp. 14393-14422, 2023.

[15] R. S. Lee et al., "A Curated Mammography Data Set for Use in Computer-aided Detection and Diagnosis Research," Scientific Data, vol. 4, p. 170177, DOI: 10.1038/sdata.2017.177, 2017.

[16] I. C. Moreira et al., "INbreast: Toward a Full-field Digital Mammographic Database," Academic Radiology, vol. 19, no. 2, pp. 236– 248, DOI: 10.1016/j.acra.2011.09.014, 2012.

[17] H. T. Nguyen et al., "VinDr-Mammo: A Large-scale Benchmark Dataset for Computer-aided Diagnosis in Full-field Digital Mammography," Scientific Data, vol. 10, p. 277, DOI: 10.1038/s41597-023- 02100-7, 2023.

[18] G. I. Quintana et al., "Exploiting Patch Sizes and Resolutions for Multi-scale Deep Learning in Mammogram Image Classification," Bioengineering, vol. 10, no. 5, p. 534, DOI: 10.3390/bioengineering10050534, 2023.

[19] P. Oza, P. Sharma, S. Patel, F. Adedoyin and A. Bruno, "Image Augmentation Techniques for Mammogram Analysis," Journal of Imaging, vol. 8, no. 5, p. 141, DOI: 10.3390/jimaging8050141, 2022.

[20] A. P. Adedigba, S. A. Adeshina and A. M. Aibinu, "Performance Evaluation of Deep Learning Models on Mammogram Classification Using Small Dataset," Bioengineering, vol. 9, no. 4, p. 161, 2022.

[21] A. G. M. Al Mansour, F. Alshomrani, A. Alfahaid and A. T. M. Almutairi, "MammoViT: A Custom Vision Transformer Architecture for Accurate BI-RADS Classification in Mammogram Analysis," Diagnostics, vol. 15, no. 3, p. 285, DOI: 10.3390/diagnostics15030285, 2025.

[22] S. Elumalai et al., "Breast Cancer Classification Based on Micro-calcifications Using Dual Branch Vision Transformer Fusion," Scientific Reports, vol. 15, DOI: 10.1038/s41598- 025-34377-6, 2025.

[23] S. Sharma, Y. Singh and T. Choudhury, "Advanced Deep Learning Architectures for Enhanced Mammography Classification: A Comparative Study of CNNs and ViT," Discover Artificial Intelligence, vol. 5, Article no. 187, DOI: 10.1007/s44163-025-00426-2, 2025.

[24] S. Ghosh et al., "Mammo-CLIP: A Vision Language Foundation Model to Enhance Data Efficiency and Robustness in Mammography," Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI 2024), vol. LNCS 15012, DOI: 10.1007/978-3-031-72390-2_59, Morocco, 2024.

[25] F. M. Talaat et al., "Grad-CAM Enabled Breast Cancer Classification with a 3D Inception-ResNet V2: Empowering Radiologists with Explainable Insights," Cancers, vol. 16, no. 21, p. 3668, 2024.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks *via* Gradient-based Localization," Proc. of the IEEE Int. Conf. on Computer Vision (ICCV), pp. 618–626, Venice, Italy, 2017.

[27] N. Shifa, M. Saleh, Y. Akbari and S. Al Maadeed, "A Review of Explainable AI Techniques and Their Evaluation in Mammography for Breast Cancer Screening," Clinical Imaging, vol. 123, p. 110492, 2025.

[28] H. C. Lu, E. W. Loh and S. C. Huang, "The Classification of Mammogram Using Convolutional Neural Network with Specific Image Pre-processing for Breast Cancer Detection," Proc. of the 2nd IEEE Int. Conf. on Artificial Intelligence and Big Data (ICAIBD), pp. 9–12, Chengdu, China, 2019.

[29] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Proc. of the 36th Int. Conf. on Machine Learning (ICML), pp. 6105–6114, PMLR 97, L. B., California, USA, 2019.

[30] S. R. Salian and S. D. Sawarkar, "Melanoma Skin Lesion Classification Using Improved Efficient-NetB3," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 8, no. 1, pp. 45–56, DOI: 10.5455/jjcit.71-1636005929, 2022.

[31] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[32] G. Huang et al., "Densely Connected Convolutional Networks," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, 2017.

[33] Z. Liu et al., "A ConvNet for the 2020s," Proc. of the 2022 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986, 2022.

[34] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv Preprint, arXiv: 2010.11929, 2021.

[35] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," Proc. of the 2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV), pp. 9992–10002, 2021.

[36] E. R. DeLong et al., "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.

[37] X. Sun and W. Xu, "Fast Implementation of DeLong's Algorithm for Comparing the Areas under Correlated Receiver Operating Characteristic Curves," IEEE Signal Process. Lett., vol. 21, no. 11, pp. 1389–1393, 2014.

[38] C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks," Proceedings of the 34th International Conference on Machine Learning (ICML'17) – vol. 70, pp. 1321 – 1330, 2017.

[39] W. J. Youden, "Index for Rating Diagnostic Tests," Cancer, vol. 3, no. 1, pp. 32–35, 1950.

[40] W. M. Salama and M. H. Aly, "Deep Learning in Mammography Images Segmentation and Classification: Automated CNN Approach," Alexandria Engineering Journal, vol. 60, no. 5, pp. 4701–4709, 2021.

[41] Z. Jafari and E. Karami, "Breast Cancer Detection in Mammography Images: A CNN-based Approach with Feature Selection," Information, vol. 14, no. 7, DOI: 10.3390/info14070410, 2023.

ملخص البحث:

غالباً ما تكون معدّلات الدّقّة المرتفعة المبلَّغ عنها في مجموعات بيانات تصوير الثّدي بالأشعة السّينية مبالغاً فيها بسبب تسرُّب البيانات الذي ينجم غالباً عن عمليات تعزيز البيانات المطبّقة قبل تقسيم البيانات إلى مجموعاتٍ للتّدريب والاختبار. فعلى سبيل المثال، يؤدّي مجرّد تغيير ترتيب التّقسيم في مجموعة بيانات (Mini-MIAS) إلى انخفاض دقة نموذج يعتمد على استخراج السّمات العميقة وآلة المتّجهات الدّاعمة من 0.976 إلى 0.513، وهي نسبة تعادل الصّدفة العشوائية، في حين تظلّ الإشارة التّنبؤية صالحة عند مستوى رقعة الآفة.

تأتي مساهمتنا على مستوى البروتوكول المنهجي وليس على مستوى البنية المعمارية للنّموذج، وتتمثّل في إجراء تدقيقٍ قابل للتّكرار لكشف تسرُّب البيانات، وتحديد ومعالجة مشكلة تداخل المرضى، حيث تتبّعنا السّبب إلى استخلاص أقسام الكُتل الورمية والتكلّسات بشكلٍ مستتلّ، مما شمل 31 مريضاً؛ أي 8.5% من آفاق مجموعة الاختبار في مجموعة البيانات (CBIS-DDSM)، وإجراء مسح شامل للبنى المعمارية لاختبار سقف الأداء بدلاً من افتراضه، وابتكار إجراء لا يعتمد على التّصنيفات لنقل عتبة اتّخاذ القرار إلى مؤسسة طبية جديدة.

وقد حقّق نموذج خفيف الوزن يعتمد على معالج الرُّقع تمّ تدريبه على مجموعة بيانات (CBIS-DDSM) وفق بروتوكول صارم نتائج جيدة في الفصل بين المرضى وقيماً جيدة من حيث مؤشرات الأداء اتسمت بالاستقرار لعددٍ من مجموعات البيانات التي استخدمت لتقييم أداء النموذج.

مما سبق نستنتج أنّ التّحكُّم في تسرُّب البيانات والتحقق الخارجي من صحة النّموذج هما أمران أكثر من مجرّد النّظر إلى أرقام الدّقّة المعلنة عند التّعامل مع مجموعات بيانات آفات الثّدي.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 03, September 2026.

FILTER SPECTRAL EFFICIENCY (FSE): STRUCTURED NETWORK PRUNING *VIA* INTER-LAYER SPECTRAL COUPLING

Soumia Laroui[1], Hamida Seba[2], Karima Amrouche[1] and Yassine Ait Ali Yahia[1]

(Received: 12-Jul.-2026, Revised: 2-Sep.-2026 and 4-Sep.-2026, Accepted: 4-Sep.-2026)

ABSTRACT

Structured filter pruning is an effective approach for reducing the computational cost of deep neural networks while preserving efficient deployment on standard hardware. However, most existing pruning methods evaluate filters independently, overlooking the inter-layer spectral coupling between consecutive layers, while many high-performing approaches rely on activation statistics or calibration data. We propose Filter Spectral Efficiency (FSE), a data-free structured pruning framework that estimates filter importance through the spectral coupling between adjacent layers. FSE constructs singular-value-based filter descriptors, builds inter-layer coupling matrices, derives a spectral pruning criterion and allocates sparsity through a global adaptive budget. Extensive experiments on five CNN architectures for CIFAR-10 and ResNet-50 on ImageNet demonstrate that FSE consistently achieves a favorable accuracy-compression trade-off, outperforming representative structured pruning methods at comparable parameter and FLOP reductions. On CIFAR-10, FSE reduces the FLOPs of ResNet-56 by 59.8% with only a 0.07 percentage point accuracy drop and prunes GoogLeNet by 70.7% FLOPs while improving Top-1 accuracy over H Rank by 0.73 percentage points. On ImageNet, FSE improves the Top-1 accuracy of ResNet-50 by 1.14 percentage points over H Rank at a comparable FLOP reduction. These results demonstrate that inter-layer spectral coupling provides an effective and practical foundation for data-free structured network pruning.

KEYWORDS

Structured pruning, Data-free pruning, Inter-layer spectral coupling, Singular value decomposition, Deep neural networks.

1. INTRODUCTION

Deep neural networks (DNNs) have achieved remarkable success in visual-recognition tasks [1]-[2], but their high computational and memory demands hinder deployment on resource-constrained platforms such as mobile and embedded devices. Modern convolutional architectures, including ResNet [3], DenseNet [4] and GoogLeNet [5], contain millions of parameters and require billions of floating-point operations (FLOPs) per inference. Consequently, reducing computational cost while preserving predictive performance has become a central challenge in deep learning [6]. Beyond deployment efficiency, the increasing computational demands of deep learning have raised concerns about energy consumption and environmental impact. Green AI [7] advocates evaluating computational efficiency alongside accuracy, making model compression important for both resource-constrained deployment and sustainable AI.

Among model compression techniques [8]-[11], structured filter pruning has emerged as one of the most practical and hardware-friendly approaches, producing compact dense networks with genuine reductions in parameters, FLOPs, memory footprint and inference latency while remaining directly deployable on standard hardware [6], unlike unstructured pruning, which typically requires specialized sparse software or hardware to realize comparable acceleration [12]. This naturally raises the central question: Which filters can be removed with the least impact on the network's predictive performance?

Weight-based and data-dependent pruning criteria have been pursued in parallel rather than in succession. Weight-based methods estimate filter importance directly from the learned weights [13]-[14], whereas data-dependent methods exploit activation responses or calibration data [15]-[16]; both lines have continued to develop in parallel in subsequent work [17]-[19]. The data-free setting is nevertheless of practical importance, since data-dependent scoring adds computational cost and

1. S. Laroui, K. Amrouche and Y. Ait Ali Yahia are with Laboratoire de la Communication dans les Systèmes Informatiques (LCSI), Ecole nationale Supérieure d'Informatique (ESI), BP, 68M Oued-Smar, 16270 Alger, Algeria. Emails: {s_laroui; k_amrouche; y_ait_ali_yahia}@esi.dz
2. H. Seba is with Universite Claude Bernard Lyon 1, CNRS, INSA Lyon, LIRIS, Villeurbanne, France. Email: hamida.seba@univ-lyon1.fr

representative data may be unavailable for privacy, security or proprietary reasons. Independently of this distinction, many existing criteria assess filters primarily through individual properties or relationships within the same layer, without explicitly characterizing how the output of one layer couples to the representation consumed by the following layer. This leads to the following question: Can filter importance be estimated entirely from the network weights while explicitly modeling inter-layer spectral coupling between consecutive layers?

We hypothesize that filter importance is influenced not only by properties of individual filters and within layer redundancy, but also by the spectral coupling between a filter's output representation and those of the following layer. In other words, a filter may matter not merely because of its own magnitude or within-layer redundancy, but also because of how effectively its output interacts with the following layer. This hypothesis is tested in Sub-section 4.3: at the same FLOP reduction and using the same singular-value descriptors, replacing the coupling-based cost with a within-layer spectral score reduces Top-1 accuracy by 0.36 percentage points on ResNet-56/CIFAR-10, providing empirical support for the relevance of inter-layer spectral coupling to filter importance.

Motivated by this hypothesis, we introduce Filter Spectral Efficiency (FSE), a structured pruning framework with entirely data-independent filter-importance estimation. FSE first derives SVD-based spectral descriptors directly from the learned weights, then constructs an inter-layer coupling matrix the singular spectrum of which characterizes the spectral coupling between consecutive layers. From this spectrum, FSE estimates a per-filter spectral cost and integrates these costs into a global adaptive pruning budget that ranks filters across the network according to redundancy rather than fixed per-layer pruning ratios. Consequently, FSE prunes aggressively in redundant regions while preserving spectrally critical ones. All pruning decisions are made before any data is seen; only the subsequent fine-tuning uses training data, following standard structured pruning practice.

We evaluate FSE on CIFAR-10 [20] using representative CNN architectures spanning the dominant design paradigms: VGG-16 [21], ResNet-56/110 [3], DenseNet-40 [4] and GoogLeNet [5]. We further validate its scalability on the ImageNet ILSVRC-2012 benchmark [22] with ResNet-50. Across all settings, FSE achieves a favorable accuracy-compression trade-off, matching or outperforming recent structured-pruning methods, such as HRank [15], CHIP [23] and the Shapley-value criterion [17], while delivering genuine reductions in computational cost.

The main contributions of this paper are:

- A relational spectral filter-importance criterion. We propose an inter-layer spectral coupling criterion that estimates filter importance directly from the learned weights by modeling information transfer between consecutive layers.
- A global adaptive pruning strategy. FSE ranks filters network-wide according to spectral cost and allocates pruning adaptively under a single sparsity target, avoiding fixed layer-wise pruning ratios.
- A data-independent (data-free) pruning framework. All pruning decisions are derived directly from the learned weights, requiring no calibration data, activation statistics or feature-map collection.
- Comprehensive validation. Experiments on five CNN architectures for CIFAR-10 and ResNet-50 on ImageNet demonstrate that FSE matches or surpasses strong structured-pruning baselines while delivering genuine reductions in parameters and FLOPs.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed Filter Spectral Efficiency (FSE) framework, including the inter-layer coupling formulation, spectral cost, global adaptive pruning budget and the structural surgery stage for dependency-aware network reconstruction. Section 4 describes the experimental setup and presents comparisons with state-of-the-art methods together with ablation studies. Finally, Section 5 concludes the paper and discusses future-research directions.

2. RELATED WORK

Structured pruning removes coherent computational units, such as filters, channels or blocks, yielding smaller dense networks that accelerate directly on standard hardware, unlike unstructured pruning [12], the irregular sparsity of which often requires dedicated sparse kernels [6], [13]. FSE belongs to this family and evaluates compression through the actual reduction in FLOPs and parameters. We position

FSE along two axes: how filter importance is estimated across layers and whether the pruning decision requires data.

2.1 Modeling Cross-layer Interactions in Filter Pruning

Most existing criteria assess filters from individual properties or relationships within a layer. Magnitude-based methods [13] use ℓ_1/ℓ_2 norms, while FPGM [14] identifies within-layer geometric redundancy. Data-dependent methods, such as H Rank [15], CHIP [23] and FilterSketch [24], exploit activation or reconstruction information collected one layer at a time [25]-[26]. Learning-based approaches, such as GAL [16], DECORE [27], Auto- ℓ_0 pruning [18] and related methods [19], [28], optimize filter or layer gates, while explainability-based methods, such as Shapley-value pruning [17], estimate marginal contributions through attribution or sampling. These approaches generally do not explicitly model dependencies between consecutive layers. CIE [29] instead estimates cross-layer importance empirically through per-layer pruning and fine-tuning trials and uses the resulting contribution information to determine layer-wise sparsity. Thus, CIE incorporates cross-layer information through empirical pruning responses rather than through an explicit spectral coupling derived from the learned weights. FSE instead defines filter importance through the inter-layer spectral coupling between consecutive convolutional representations, computed analytically from the learned weights without activation sampling, attribution estimation or pruning trials.

2.2 Data-free Pruning

Beyond the cross-layer distinction, pruning methods also differ in whether their pruning decisions require data. Data-dependent criteria, such as HRank [15], CHIP [23] and Shapley-value pruning [17], require representative data to collect activations or evaluate filter contributions, introducing computational overhead and potential privacy or data-availability constraints. Data-free methods instead derive importance directly from the learned weights. Representative examples include magnitude pruning [13], FPGM [14] and Soft Filter Pruning [30]. CORING [31] applies higher-order SVD to measure within-layer filter similarity, while RASP [32] derives amplitude-based saliency from learned parameters and uses layer-sensitive regularization for global filter selection. However, removing data dependence from the importance criterion does not necessarily determine how sparsity should be allocated across layers. Fixed ratios, sensitivity analysis or additional optimization/search may still be required, as in Network Slimming [33], ABCPruner [34], AFPruner [35] and Auto- ℓ_0 Pruner [18]. FSE is data-independent throughout the pruning-decision stage: it derives filter importance from SVD-based descriptors of the learned weights and their inter-layer coupling and allocates the global sparsity budget directly from the resulting spectral cost, with only lightweight per-layer safety constraints. No calibration data, activation statistics, forward passes or auxiliary search/optimization are required to make the pruning decision; training data is used only in the subsequent fine-tuning stage.

2.3 Positioning of This Work

Despite their methodological diversity, the reviewed criteria differ in how they incorporate information beyond an individual filter. Most methods rely on within-layer properties or pairwise relationships, while a smaller body of work incorporates cross-layer information through empirical pruning responses. FSE departs from these approaches by using the spectral relationship between consecutive convolutional representations as the filter-importance criterion, while remaining fully data-independent at the pruning-decision stage. It differs from existing structured-pruning methods in three fundamental aspects:

1) an inter-layer spectral criterion that models the coupling between consecutive convolutional representations, providing a different formulation from existing cross-layer approaches;
2) a fully data-independent (data-free) pruning-decision mechanism that derives the required spectral information solely from the learned weights, without calibration data or activation statistics; and
3) an integrated global pruning strategy that combines the proposed spectral cost with adaptive network-wide sparsity allocation, rather than treating filter importance and pruning-budget allocation as separate steps.

Hybrid compression pipelines combine structured pruning with complementary techniques. NORTON [36] applies tensor decomposition followed by pruning using a principal-angle-based sub-space criterion, whereas CoDeC [37] clusters filters for coupled tensor decomposition using an SVD-based

sub-space measure. Both approaches exploit singular sub-spaces, but differ from FSE in scope and objective: NORTON prunes decomposition factors after tensor decomposition, CoDeC clusters filters within layers, whereas FSE uses rank-one SVD descriptors to model inter-layer coupling before structural pruning. Applying FSE as a pre-processing stage could reduce both channel dimensions before factorization; moreover, the remaining singular values obtained during the SVD, although discarded by the current FSE descriptor, could guide subsequent rank selection through their spectral decay. STPP [38] combines structured pruning with knowledge distillation; for FSE, the original unpruned network is already available as a teacher, making post-pruning feature-level distillation a natural extension. These combinations are left for future work.



Figure 1. Overview of the FSE pruning pipeline.

The following section presents the proposed Filter Spectral Efficiency (FSE) framework and details the formulation of the inter-layer coupling criterion, the global adaptive pruning budget and the overall pruning algorithm.

3. METHODOLOGY

Most structured-pruning criteria score each filter in isolation; by its ℓ_1 magnitude, its batch-normalization scale or its first-order effect on the loss, capturing how large a filter is, but not how its role relates to the following layer. Yet the output channels of one convolution are exactly the input channels of the next, so a filter's practical redundancy depends as much on this inter-layer relationship as on its own magnitude. Filter Spectral Efficiency (FSE) exploits this observation: it represents each filter by a compact spectral descriptor, measures how the descriptors of two consecutive layers couple and prunes the filters that contribute least to that coupling. FSE is data-free; operating purely on the learned weights, it yields a genuinely structured model compatible with standard inference devices.

The pipeline comprises five stages:

1) a per-filter spectral descriptor (Sub-section 3.1);
2) an inter-layer coupling matrix and its spectrum (Sub-section 3.2);
3) a per-filter spectral pruning cost (Sub-section 3.3);
4) a global spectral budget (Sub-section 3.4); and
5) the structural surgery that removes the selected channels (Sub-section 3.5).

The procedure is summarized in Algorithm 1 and illustrated in Figures 1 and 2-4. Notation. We consider a convolutional network with L prunable convolutional layers indexed by $\in \{0, \dots, L-1\}$. The weight tensor of layer l is

$$\boldsymbol{W}^{(l)} \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{in}}^{(l)} \times k \times k}, \tag{1}$$

where $C_{\text{out}}^{(l)}$ is the number of filters, $C_{\text{in}}^{(l)}$ the number of input channels and k the kernel size; the i th filter is $\boldsymbol{W}_i^{(l)} \in \mathbb{R}^{C_{\text{in}}^{(l)} \times k \times k}$. Each convolution is followed by a batch-normalization (BN) layer the parameters of which $\boldsymbol{\theta}_{\text{BN}}^{(l)} = \left(\boldsymbol{\gamma}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\mu}^{(l)}, \boldsymbol{v}^{(l)}\right)$ are pruned jointly with the corresponding filters. Finally, we define the protected set of layers

$$\mathcal{S} = \mathcal{S}_{\text{skip}} \cup \{0, L-1\}, \tag{2}$$

where $\mathcal{S}_{\text{skip}} \subseteq \{0, \dots, L-1\}$ is the sub-set of layers automatically detected as skip-connection (shortcut) layers during the Conv-BN parsing stage of Algorithm 1 and the indices 0 and $L-1$ denote the first and the last convolution, respectively. By convention we set $\kappa_i^{(l)} = +\infty$ for every i and every $l \in \mathcal{S}$, so that the output channels of a protected layer are never removed by the budget allocation of Eqs. (13)-(15); this preserves tensor-shape compatibility across residual branches and at the network boundaries.

3.1 Spectral Filter Descriptor

Existing weight-based pruning criteria commonly represent a filter by a scalar magnitude or other aggregate statistic, which may overlook the dominant structure encoded across its input channels. Since FSE aims to compare filters across layers using their learned weights, we require a compact representation that preserves the most informative spectral structure of each filter. We therefore use the leading singular component of each filter as its spectral descriptor, providing a data-independent representation of its dominant input-channel structure. Each filter is summarized by a compact vector capturing its dominant behaviour across input channels. The spatial dimensions are unfolded into a matrix

$$\boldsymbol{F}_i^{(l)} = \text{reshape}\left(\boldsymbol{W}_i^{(l)}\right) \in \mathbb{R}^{C_{\text{in}}^{(l)} \times k^2}, \tag{3}$$

where each row is the spatial weighting pattern of one input channel. Its singular value decomposition (SVD)

$$\boldsymbol{F}_i^{(l)} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\top}, \tag{4}$$

has singular values $s_1 \geq s_2 \geq \cdots$ with left singular vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots$. The leading singular triplet $(s_1, \boldsymbol{u}_1, \boldsymbol{v}_1)$ provides the best rank-one approximation of the filter, so we retain only this dominant component:

$$\boldsymbol{c}_i^{(l)} = s_1 \boldsymbol{u}_1 \in \mathbb{R}^{C_{\text{in}}^{(l)}}. \tag{5}$$

Scaling the dominant direction $\boldsymbol{u}_1$ by s_1 preserves both the principal input-channel structure and its strength. Since the SVD fixes $\boldsymbol{u}_1$ only up to sign, we enforce the largest-magnitude entry to be positive,

$$\boldsymbol{c}_i^{(l)} \leftarrow \text{sgn}\left(c_{i,j^\star}^{(l)}\right) \boldsymbol{c}_i^{(l)}, j^\star = \arg\max_j \left|c_{i,j}^{(l)}\right|, \tag{6}$$

ensuring a deterministic representation for subsequent comparisons. This convention is a matter of bookkeeping rather than of geometry: the sign returned by an SVD routine is implementation-dependent, whereas the pruning cost of Eq. (12) is invariant to any per-filter sign choice (Remark 1). Hence, no preferred orientation is imposed on the dominant singular direction and the resulting pruning criterion is independent of the arbitrary sign convention.

$$\boldsymbol{D}^{(l)} = \left[\boldsymbol{c}_0^{(l)}, \ldots, \boldsymbol{c}_{C_{\text{out}}^{(l)}-1}^{(l)}\right]^{\top} \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{in}}^{(l)}}. \tag{7}$$

Each row of $\boldsymbol{D}^{(l)}$ acts as a fingerprint of a filter; filters with similar fingerprints capture similar dominant structures and may indicate functional redundancy (Figure 2).

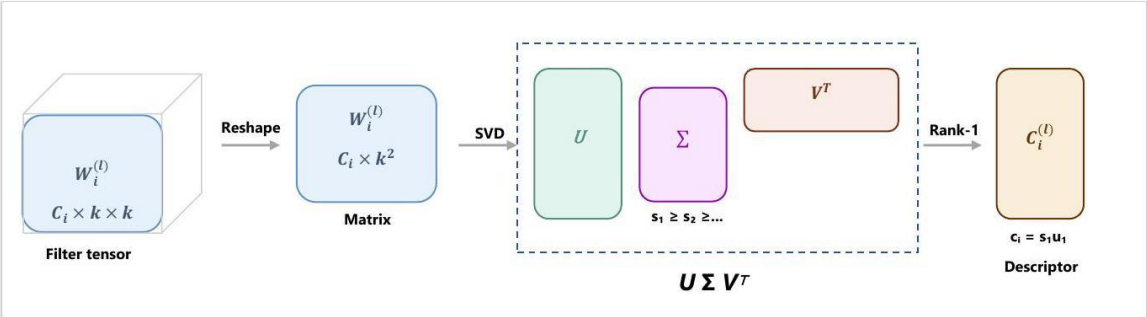


Figure 2. Construction of the per-filter spectral descriptor.

3.2 Inter-layer Coupling Matrix and Spectrum

A filter descriptor alone captures only the intrinsic structure of an individual filter and therefore cannot reveal how that filter interacts with the representation processed by the following layer. To explicitly capture this dependency while remaining data-independent, FSE compares the spectral descriptors of consecutive layers in a common representation space. We therefore construct an inter-layer coupling matrix based on cosine similarity and analyze its singular spectrum to characterize the dominant patterns of cross-layer alignment. First, when the descriptor dimensionality is large, descriptors are projected onto their leading principal directions to discard low-variance (noise-prone) components and reduce cost. Let $\boldsymbol{D}^{(l)\top} = \boldsymbol{U}_D \boldsymbol{\Sigma}_D \boldsymbol{V}_D^{\top}$ and let $d_{\max} = 16$ be the maximum retained dimension. The reduced

descriptor matrix is

$$\widetilde{\boldsymbol{D}}^{(l)} = \begin{cases} \boldsymbol{D}^{(l)}\boldsymbol{U}_D[:,1:d_{\max}], & C_{\text{in}}^{(l)} > d_{\max}, \\ \boldsymbol{D}^{(l)}, & \text{otherwise}. \end{cases} \tag{8}$$

Second, descriptors of adjacent layers may differ in dimensionality. Rather than truncating, the lower-dimensional descriptors are zero-padded to $d^\star = \max(d_1, d_2)$; since cosine similarity depends only on orientation, appending zeros leaves all pairwise similarities unchanged:

$$\widetilde{D} \mapsto \left[\widetilde{D} \| \mathbf{0}_{C_{\text{out}} \times (d^\star - d)}\right]. \tag{9}$$

Coupling is then measured through pairwise cosine similarities, appropriate here, because descriptors are compared by dominant direction rather than magnitude. Let $\widehat{\boldsymbol{D}}^{(l)}$ denote the ℓ_2-normalized rows of the padded $\widetilde{\boldsymbol{D}}^{(l)}, \widehat{\boldsymbol{d}}_\iota = \boldsymbol{d}_i / \|\boldsymbol{d}_i\|_2$. The coupling matrix is

$$\boldsymbol{M}^{(l)} = \widehat{\boldsymbol{D}}^{(l)}\left(\widehat{\boldsymbol{D}}^{(l+1)}\right)^\top \in \mathbb{R}^{C_{\text{out}}^{(l)} \times C_{\text{out}}^{(l+1)}}, M_{ij}^{(l)} = \cos\left(\boldsymbol{d}_i^{(l)}, \boldsymbol{d}_j^{(l+1)}\right), \tag{10}$$

where large values indicate that filter i in layer l and filter j in layer $l+1$ are strongly aligned. For a sequential pair of convolutional representations, Eq. (10) defines the pairwise spectral coupling used for filter-importance estimation. It is not intended to encode the complete computational dependency graph of the network. For architectures with non-sequential connectivity, such as DenseNet-40 and GoogLeNet, the FSE criterion retains this pairwise spectral formulation, while the subsequent structural-pruning stage uses DepGraph [39] to trace dependencies induced by dense concatenation, parallel branches and downstream operations and to propagate the selected channel-removal decisions accordingly.

The global structure of this alignment is revealed by the SVD

$$\boldsymbol{M}^{(l)} = \boldsymbol{U}^{(l)}\boldsymbol{\Sigma}^{(l)}\boldsymbol{V}^{(l)\top}, \sigma_1^{(l)} \geq \cdots \geq \sigma_r^{(l)} \geq 0, \tag{11}$$

the singular values $\left\{\sigma_k^{(l)}\right\}$ of which form the coupling spectrum. Each $\sigma_k^{(l)}$ measures the strength of an independent-alignment pattern, so the spectrum characterises the effective dimensionality of the inter-layer coupling. When spectral energy concentrates in few singular values, the coupling is effectively low-rank: many filters share the same dominant interaction patterns, indicating substantial redundancy.

3.3 Per-filter Pruning Cost

The coupling spectrum describes the importance of the different inter-layer alignment modes, but pruning requires converting this global spectral information into a filter-level quantity. A direct filter-level cost is therefore needed to rank filters according to their contribution to the cross-layer interaction. FSE obtains this cost from the participation of each filter in the singular modes, weighted by the corresponding singular values. A filter that contributes little to the spectrum carries little of the information exchanged between the two layers and is a pruning candidate, whereas a strongly contributing filter is structurally important. Reusing the SVD of (11), we define the pruning cost of filter i as z

$$\kappa_i^{(l)} = \sum_k \left(U_{ik}^{(l)}\right)^2 \sigma_k^{(l)}. \tag{12}$$

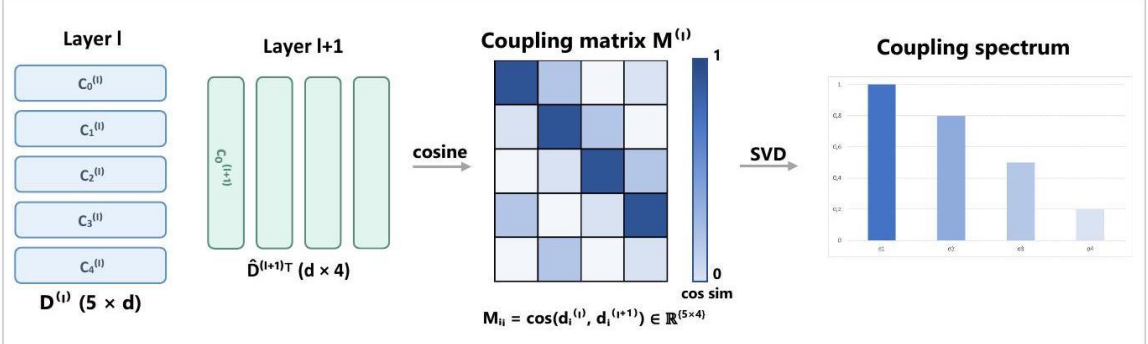


Figure 3. A representative inter-layer coupling matrix and its spectrum.

Here, $\left(U_{ik}^{(l)}\right)^2$ measures the participation of filter i in spectral mode k and $\sigma_k^{(l)}$ weights that mode by its significance, so filters participating strongly in dominant modes receive a large cost and are preferentially retained. The cost requires only $\mathcal{O}(C)$ per-layer computation, yet Proposition 1 shows it exactly decomposes the nuclear norm of the coupling matrix across filters, providing a principled rather than heuristic justification.

Protected filters. For any protected layer $l \in \mathcal{S}$ and for the final layer where no forward coupling is defined, all filters receive infinite cost, $\kappa_i^{(l)} = +\infty$, guaranteeing that they are never removed while keeping a uniform framework across layers.

3.4 Global Spectral Budget

A filter-level importance score does not by itself determine how many filters should be removed from each layer. Fixed layer-wise ratios impose the same pruning pressure on layers with potentially different levels of spectral redundancy, while separate sensitivity or search procedures introduce additional computational cost. Since FSE already provides a spectral cost for every filter, we use these costs to allocate a single global pruning budget across the network, allowing the pruning profile to emerge from the measured spectral redundancy. Let $\mathcal{E} = \left\{(l,i): \kappa_i^{(l)} < \infty\right\}$ be the eligible filters and $s \in [0,1]$ the target sparsity. The number of filters to remove is

$$T = \lfloor s|\mathcal{E}| \rfloor. \tag{13}$$

To prevent collapsing a layer, each layer l (with n_l eligible filters) is capped:

$$\text{cap}_l = \begin{cases} 0, & n_l \le 1, \\ \max\left(0, \min\left(n_l - 1, \lfloor n_l \rho_{\max} \rfloor\right)\right), & n_l > 1, \end{cases} \tag{14}$$

where $\rho_{\max}$ (default 0.8) is the maximum removable fraction per layer; this keeps at least one filter per non-trivial layer and avoids over-concentrated pruning. Eligible filters are ranked by ascending cost and selected greedily until the budget is exhausted, subject to the caps:

$$\mathcal{P} = \left\{(l,i): (l,i) \text{ visited in ascending order of } \kappa_i^{(l)}, |\mathcal{P}| \le T, |\{(l,\cdot) \in \mathcal{P}\}| \le \text{cap}_l\right\}. \tag{15}$$

The effective sparsity is $|\mathcal{P}|/|\mathcal{E}|$; if the caps prevent the target from being reached, the shortfall is reported so the user may relax $\rho_{\max}$ or lower s. Because ranking is network-wide, FSE removes more filters from spectrally redundant layers and fewer from important ones, so the allocation emerges from the spectral costs rather than manual ratios (Figure 4). The greedy rule is also optimal: since costs are additive and feasibility is encoded by the caps, selecting the lowest-cost feasible filters minimizes the total pruning cost over all feasible selections of equal cardinality (Proposition 2).

3.5 Structural Surgery

Selecting filters is not sufficient to obtain a valid compact network, because removing an output channel from one layer can change the input-channel structure of downstream layers. This issue is particularly important for architectures with complex connectivity, such as dense concatenations and parallel branches. FSE therefore performs dependency-aware structural surgery to preserve the computational graph after pruning; for DenseNet-40 and GoogLeNet, DepGraph [39] is used to trace the resulting channel dependencies and propagate the pruning decisions to affected layers.

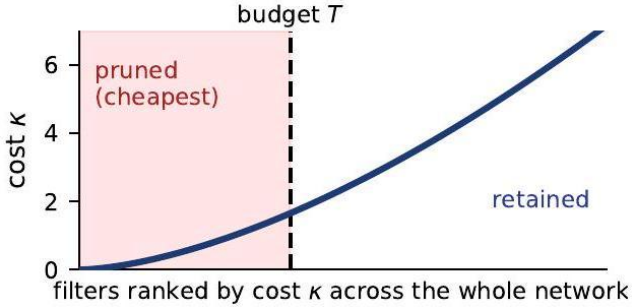


Figure 4. Global cost-ranked filter budget.

Algorithm 1. Filter Spectral Efficiency (FSE) pruning.

Require: network f; target sparsity s; cap $\rho_{\max}$
extract Conv-BN pairs and build the protected set $\mathcal{S}$
for each layer l **do**
  compute descriptors $\boldsymbol{D}^{(l)}$ via Eqs. (3)-(7)
end for
for $l = 0$ to $L - 2$ **do**
  form $\boldsymbol{M}^{(l)}$ (Eq. (10)) and its SVD (Eq. (11))
end for
for each layer l **do**
  if $l \in \mathcal{S}$ or coupling undefined **then**
    $\kappa_i^{(l)} \leftarrow \infty \forall i$
  else
    compute $\kappa^{(l)}$ via Eq. (12)
  end if
end for
select $\mathcal{P}$ by the global budget (Eqs. (13)-(15))
perform structural surgery (Eqs. (16)-(18))
return pruned network (After fine-tuned)

The final stage turns the pruning set $\mathcal{P}$ into a physically smaller network. For each layer l, let $\mathcal{P}_l = \{i: (l,i) \in \mathcal{P}\}$ be the pruned output channels and $\mathcal{K}_l = \left\{0, \dots, C_{\text{out}}^{(l)} - 1\right\} \setminus \mathcal{P}_l$ the retained ones. The pruned channels are removed from the kernel and BN parameters,

$$\boldsymbol{W}^{(l)} \leftarrow \boldsymbol{W}^{(l)}[\mathcal{K}_l, :, :, :], \tag{16}$$

$$\boldsymbol{\theta}_{\text{BN}}^{(l)} \leftarrow \boldsymbol{\theta}_{\text{BN}}^{(l)}[\mathcal{K}_l]. \tag{17}$$

and the decision is propagated to every downstream convolution g consuming layer l :

$$\boldsymbol{W}^{(g)} \leftarrow \boldsymbol{W}^{(g)}[:, \mathcal{K}_l, :, :]. \tag{18}$$

4. EXPERIMENTS

We evaluate FSE on CIFAR-10 and ImageNet using six CNN architectures, compare it with representative structured-pruning methods and conduct ablation studies on the proposed inter-layer spectral criterion and global spectral budget.

4.1 Experimental Setup

Datasets. Experiments are conducted on CIFAR-10 [20] and ImageNet ILSVRC-2012 [22]. CIFAR-10 consists of 60,000 color images of size 32 × 32 from 10 classes, split into 50,000 training and 10,000 test images. ImageNet ILSVRC-2012 contains approximately 1.28 million training images and 50,000 validation images across 1,000 classes with the shape of 224 × 224.

Architectures. On CIFAR-10, we evaluate FSE on ResNet-56, ResNet-110 [3], DenseNet-40 [4], GoogLeNet [5] and VGG-16 [21], covering residual, densely connected, inception-style and plain CNN architectures. For large-scale evaluation, we also prune ResNet-50 [3] on ImageNet.

Implementation details. FSE operates directly on trained weights without calibration data. On CIFAR-10, pruned models are fine-tuned for 160 epochs using SGD (momentum 0.9, weight decay 5×10^{-4}), an initial learning rate of 0.01 with cosine decay and a batch size of 128. On ImageNet, ResNet-50 is finetuned for 90 epochs using SGD with the same momentum and weight decay, an initial learning rate of 0.01 with cosine annealing, a batch size of 256 and AMP. Unless stated otherwise, FSE uses the proposed pruning criterion, a descriptor cap $d_{\max} = 16$ and a maximum per-layer pruning ratio $\rho_{\max} = 0.8$.

Metrics and protocol. On CIFAR-10, we report top-1 accuracy together with parameter and FLOP reductions relative to the unpruned baseline. On ImageNet, we report top-1, top-5 accuracy, parameter reduction and FLOP reduction. For baselines, we report results from the original papers when available. A dash "-" indicates that the corresponding metric was not reported. We ensure that comparisons use

the same network architectures.

Table 1. Comparison with state-of-the-art pruning methods on VGG-16/CIFAR-10.

Model	Method	Top-1 Acc. ↑ / ↓ (%)	Params (M)	Params ↓ (%)	FLOPs (M)	FLOPs ↓ (%)
VGG-16	Baseline	93.02 (0.00)	14.73	0.00	314.31	0.00
	L1 [13]	93.40 (+0.38↑)	5.40	64.00	206.00	34.20
	FPGM [14]	94.00 (+0.42↑)	-	-	206.00	34.20
	DECORE-500 [27]	94.02 (+0.06↑)	5.54	63.00	203.08	35.30
	FSE (Ours)	93.43 (+0.41↑)	8.14	44.72	200.88	36.09
	GAL-0.05 [16]	92.03 (-1.93↓)	3.36	77.60	189.49	39.60
	GAL-0.1 [16]	90.73 (-3.23↓)	2.67	82.20	171.89	45.20
	Network slimming [33]	93.80 (+0.14↑)	2.30	88.50	-	51.00
	FSE (Ours)	**93.20 (+0.18↑)**	**3.93**	**73.29**	**124.23**	**60.47**
	FICP [19]	93.72 (-0.24↓)	-	85.90	-	63.20
	FSE (Ours)	**93.13 (+0.11↑)**	**3.25**	**77.95**	**111.98**	**64.37**
	HRank [15]	91.23 (-2.73↓)	1.78	92.00	73.70	76.50
	Shapley [17]	93.14 (-0.82↓)	1.90	87.30	67.09	78.60
	FSE (Ours)	**92.15 (-0.87↓)**	**1.38**	**90.66**	**66.94**	**78.70**
	Auto ℓ_0 Pruner-0.5 [18]	90.55 (-2.47↓)	1.65	88.80	34.61	88.99
	FSE (Ours)	**90.65 (-2.37↓)**	**1.07**	**92.80**	**34.36**	**89.10**

Hardware and software setup. All experiments were conducted on a workstation equipped with a single NVIDIA GeForce RTX 5090 GPU with 32 GB of VRAM (CUDA 12.9), an AMD EPYC 9654 processor with 96 CPU cores, 76.5 GB of system memory and 350 GB of SSD storage. The experiments were performed under Ubuntu 24.04.4 LTS using Python 3.12.14 and PyTorch 2.11.0+cu128, with an NVIDIA driver version 610.57.04.

4.2 Comparison with State-of-the-Art

4.2.1 Results on CIFAR-10

VGG-16. As a highly over-parameterized architecture, VGG-16 is a standard benchmark for structured pruning. As summarized in Table 1, FSE achieves a favorable accuracy-efficiency trade-off across all compression regimes. At moderate compression, 36.09% FLOPs, it improves the baseline by +0.41%, matching FPGM's +0.42% while pruning more FLOPs than the ℓ_1-norm and DECORE-500. At intermediate ratios, it remains the only method staying above baseline while saving more compute: +0.18% at 60.47% FLOPs *vs.* Network Slimming's +0.14% at only 51% and 93.13% at 64.37% FLOPs *vs.* FICP's -0.24% at 63.20%. Under aggressive pruning, FSE stays competitive: at 78.70% FLOPs, it reaches 92.15%, -0.87%, exceeding H Rank by 1.68 pp and matching Shapley's -0.82% with fewer parameters, 1.38M *vs.* 1.90M, yet without any calibration data or activation statistics. At the extreme 89.10% FLOP setting, it still surpasses Auto- ℓ_0 Pruner-0.5 on all three metrics: 90.65% *vs.* 90.55% accuracy, 1.07M *vs.* 1.65M parameters. Notably, the FLOP reduction consistently exceeds the parameter reduction, indicating that FSE preferentially removes filters in computationally dominant layers, a behavior well suited to latency-bound edge deployment.

ResNet-56. ResNet-56 is a far more challenging benchmark than VGG-16 owing to its compact design and low redundancy. As reported in Table 2, FSE sustains a favorable accuracy-compression tradeoff across all ratios. At moderate compression, it reaches 93.27%, +0.01 pp, at 53.05% FLOPs, the largest reduction in its group. As the budget grows, it keeps the lead: -0.07% at 59.84% FLOPs, vs. ℓ_0 Pruner's -0.18% and 93.03% at 62.45% FLOPs, exceeding both Auto- ℓ_0 Pruner-0.3 92.96% and Direct-Indirect Filter Pruning 93.04% at higher compression. Under aggressive pruning, the gap widens: at ≈ 72% FLOPs, it attains 92.43%, beating Shapley 92.20% and CHIP 92.05%, with the largest parameter reduction, 72.47%; and at 74.48% FLOPs, it retains 92.07%, surpassing HRank by 1.35 pp with larger cuts in both parameters, 76.09% *vs.* 68.10% and FLOPs. These results demonstrate that FSE generalizes across architectures rather than exploiting architecture-specific redundancy.

Table 2. Comparison with state-of-the-art pruning methods on ResNet-56 and ResNet110/CIFAR-10.

Model	Method	Top-1 Acc. ↑/↓ (%)	Params (M)	Params ↓ (%)	FLOPs (M)	FLOPs ↓ (%)
ResNet-56	Baseline	93.26 (0.00)	0.85	0.00	127.62	0.00
	L1 [13]	93.06 (-0.20↓)	0.73	14.12	90.90	27.60
	GAL-0.6 [16]	92.98 (-0.28↓)	0.75	11.80	78.30	37.60
	SFP [30]	93.35 (-0.24↓)	-	-	59.40	52.60
	FPGM [14]	93.49 (-0.10↓)	-	-	59.40	52.60
	FSE (Ours)	**93.27 (+0.01↑)**	**0.52**	**39.50**	**59.92**	**53.05**
	ℓ_0 Pruner (63% pruned) [18]	93.08 (-0.18↓)	0.28	67.06	51.38	59.74
	FSE (Ours)	**93.19 (-0.07↓)**	**0.42**	**50.49**	**51.25**	**59.84**
	GAL-0.8 [16]	90.36 (-2.90↓)	0.29	65.90	49.99	60.20
	Direct-Indirect Filter Pruning [28]	93.04 (-0.60↓)	-	-	-	61.50
	Auto ℓ_0 Pruner-0.3 [18]	92.96 (-0.30↓)	0.31	63.53	48.47	62.02
	FSE (Ours)	**93.03 (-0.23↓)**	**0.38**	**55.15**	**47.92**	**62.45**
	CHIP [23]	92.05 (-1.21↓)	0.24	71.80	34.79	72.30
	Shapley [17]	92.20 (-1.06↓)	0.24	71.80	34.79	72.30
	FSE (Ours)	**92.43 (-0.83)**	**0.23**	**72.47**	**35.32**	**72.32**
	HRank [15]	90.72 (-2.54↓)	0.27	68.10	32.52	74.10
	FSE (Ours)	**92.07 (-1.19↓)**	**0.20**	**76.09**	**32.56**	**74.48**
ResNet-110	Baseline	93.50 (0.00)	1.73	0.00	257.09	0.00
	L1 [13]	93.30 (-0.20↓)	1.16	32.60	155.00	38.70
	SFP [30]	93.35 (+0.18↑)	-	-	152.20	40.80
	GAL-0.5 [16]	92.55 (-0.95↓)	0.95	44.80	130.20	48.50
	FPGM [14]	93.74 (+0.06↑)	-	-	121.00	52.30
	FSE (Ours)	**93.80 (+0.30↑)**	**1.04**	**39.66**	**120.28**	**53.21**
	FilterSketch [24]	93.44 (-0.06↓)	0.69	59.90	92.48	63.30
	FSE (Ours)	93.46 (-0.04↓)	0.70	59.74	88.35	65.63
	HRank [15]	92.65 (-0.85↓)	0.53	68.70	79.30	68.60
	FSE (Ours)	**93.31 (-0.19↓)**	**0.58**	**66.34**	**78.82**	**69.34**
	Shapley [17]	93.14 (-0.36↓)	0.54	68.60	72.83	71.30
	FSE (Ours)	**93.18 (-0.29↓)**	**0.52**	**69.76**	**73.13**	**71.56**
	DECORE-175 [27]	92.71 (-0.79↓)	0.35	79.60	58.37	76.90
	FSE (Ours)	**92.93 (-0.57↓)**	**0.47**	**72.86**	**58.18**	**77.37**

ResNet-110. The deeper ResNet-110 confirms these findings. At moderate compression, 53.21% FLOPs, FSE attains the highest accuracy of all methods, 93.80%, +0.30 pp, exceeding FPGM by 0.24 pp, SFP by 0.18 pp and the ℓ_1-norm criterion by 0.20 pp while pruning the most FLOPs. It stays ahead of recent feature-based methods: 93.46% at 65.63% FLOPs vs. FilterSketch, 93.44% at 63.30% FLOPs and +0.66 pp over HRank at 69.34% FLOPs. It again outperforms Shapley at $\approx$ 71.5% FLOPs in accuracy, 93.18% *vs.* 93.14%, as well as in parameters and FLOPs. Even at the most aggressive 77.37% FLOP reduction, it retains 92.93%, surpassing DECORE-175 by 0.22 pp. These results indicate that inter-layer spectral coupling provides a stable importance measure as residual networks deepen.

GoogLeNet. GoogLeNet is harder to prune than sequential or residual nets: its Inception modules concatenate multiple branches before downstream consumption, so pruning must preserve both interlayer flow and channel consistency (Table 3). At moderate compression, 66.72% FLOPs, FSE attains the highest accuracy of all methods, 94.91%, -0.14%, exceeding ABCPruner by 1.07 pp with more parameters removed, 63.91% *vs.* 60.14% and FICP by 0.16 pp at a comparable budget. The lead widens under aggressive pruning: 94.80% at 70.68% FLOPs, +0.73 pp over HRank and at the highest setting, 73.95% FLOPs, it matches Shapley within 0.03 pp, 94.66% *vs.* 94.69%, while producing a far more compact model, 1.78M *vs.* 2.10M params, 71.18% *vs.* 65.90% removed. Across all points, FSE stays within 0.39 pp of the 95.05% baseline while cutting nearly 74% of the compute, confirming that inter-layer spectral coupling remains effective even under complex multi-branch feature aggregation.

DenseNet-40. DenseNet-40 provides a particularly challenging benchmark for structured pruning,

because its dense feature reuse leaves limited redundancy. As shown in Table 4, FSE consistently achieves the best accuracy-compression trade-off across both evaluated operating points. At moderate compression, 41.01% FLOP reduction, FSE attains the highest Top-1 accuracy 94.30%, slightly outperforming HRank while operating at a marginally higher FLOP reduction. Under more aggressive pruning, 61.07% FLOP reduction, FSE further improves upon HRank and GAL-0.05, demonstrating that the proposed inter-layer spectral criterion remains effective even when redundancy is limited. Unlike VGG-16, FSE prioritizes computational efficiency over parameter reduction, achieving the largest FLOP reduction while maintaining the highest predictive performance. Overall, these results demonstrate that FSE generalizes effectively to densely connected architectures, providing the strongest overall accuracy-compression trade-off despite the limited redundancy of DenseNet-40.

Table 3. Comparison with state-of-the-art pruning methods on GoogLeNet/CIFAR-10.

Model	Method	Top-1 Acc. ↑/↓ (%)	Params (M)	Params ↓ (%)	FLOPs (B)	FLOPs ↓ (%)
GoogLeNet	Baseline	95.05 (0.00)	6.15	0.00	1.52	0.00
	L1 [13]	94.54 (-0.51↓)	3.51	42.90	1.02	32.90
	GAL-0.05 [16]	93.93 (-1.12↓)	3.12	49.30	0.94	38.20
	GAL-ApoZ [16]	92.11 (-2.94↓)	2.85	53.70	0.76	50.00
	HRank [15]	94.53 (-0.52↓)	2.74	55.40	0.69	54.90
	FICP [19]	94.75 (-0.30↓)	-	55.00	-	64.50
	ABCPruner [34]	93.84 (-0.21↓)	2.46	60.14	0.51	66.56
	FSE (Ours)	**94.91 (-0.14↓)**	**2.23**	**63.91**	**0.52**	**66.72**
	HRank [15]	94.07 (-0.98↓)	1.86	69.80	0.45	70.40
	FSE (Ours)	**94.80 (-0.25↓)**	**2.02**	**67.22**	**0.45**	**70.68**
	Shapley [17]	94.69 (-0.36↓)	2.10	65.90	0.40	73.90
	FSE (Ours)	**94.66 (-0.39↓)**	**1.78**	**71.18**	**0.40**	**73.95**

Table 4. Comparison with state-of-the-art pruning methods on DenseNet-40/CIFAR-10.

Model	Method	Top-1 Acc. ↑/↓ (%)	Params (M)	Params ↓ (%)	FLOPs (M)	FLOPs ↓ (%)
DenseNet-40	Baseline	94.81 (0.00)	1.06	0.00	292.50	0.00
	GAL-0.01 [16]	94.29 (-0.52↓)	0.67	35.60	182.92	35.30
	HRank [15]	94.24 (-0.57↓)	0.66	36.50	167.41	40.80
	FSE (Ours)	**94.30 (-0.51↓)**	**0.77**	**26.83**	**172.56**	**41.01**
	GAL-0.05 [16]	93.53 (-1.28↓)	0.45	56.70	128.11	54.70
	HRank [15]	93.68 (-1.13↓)	0.48	53.80	110.15	61.00
	FSE (Ours)	**93.79 (-1.02↓)**	**0.50**	**52.67**	**113.86**	**61.07**

4.2.2 Results on ImageNet

ResNet-50. We finally evaluate the scalability of FSE on the large-scale ImageNet benchmark using ResNet-50. For completeness, RASP [32] and CORING [31] are not included in Table 5, because their pruning protocols are not directly comparable with the single-shot pruning protocol used in this study: RASP relies on regularization-based model selection, whereas CORING employs an iterative K-shots procedure with intermediate fine-tuning. As shown in Table 5, FSE consistently achieves a favorable accuracy-compression trade-off across all compression regimes.

At moderate compression, 46.60% FLOP reduction, FSE attains the best overall trade-off, achieving 74.81% Top-1 and 92.29% Top-5 accuracy while providing the largest parameter and FLOP reductions among comparable methods. Compared with the recent Shapley-based approach, FSE delivers higher Top-1 and Top-5 accuracy together with greater compression, without requiring calibration data or activation statistics. As compression increases, FSE continues to outperform competing methods. At 57.53% FLOP reduction, it achieves the highest accuracy among methods with comparable compression, surpassing AFPruner, FPGM, ABCPruner and Auto- ℓ_0 Pruner. Under severe compression, 62.45% FLOP reduction, FSE further exceeds HRank by 1.14 percentage points in Top-1 accuracy while simultaneously achieving larger parameter and FLOP reductions. Overall, these results

demonstrate that the proposed inter-layer spectral criterion scales effectively from CIFAR-10 to ImageNet, consistently providing an excellent balance between predictive accuracy and computational efficiency through entirely data-free pruning.

Table 5. Comparison with state-of-the-art pruning methods on ILSVRC-2012 using the ResNet50 architecture.

Model	Method	Top-1 Acc. ↑/↓ (%)	Top-5 Acc. (%)	Params (M)	Params ↓ (%)	FLOPs (B)	FLOPs ↓ (%)
ResNet-50	Baseline	76.15 (0.00)	92.88	25.56	0.00	4.13	0.00
	SSS-32 [40]	74.18 (-1.94↓)	91.91 (-0.95↓)	18.60	27.06	2.82	31.05
	He et al. [26]	72.30 (-3.83↓)	90.80 (-2.06↓)	-	-	2.73	33.30
	ThiNet [25]	72.04 (-1.57↓)	90.67 (-0.69↓)	16.94	33.7	-	36.80
	SFP [30]	74.61 (-1.54↓)	92.06 (-0.81↓)	-	-	-	41.80
	DECORE-6 [27]	74.58 (-1.57↓)	92.18 (-0.69↓)	14.10	44.71	2.36	42.30
	GAL-0.5 [16]	71.95 (-4.18↓)	90.94 (-1.92↓)	21.20	16.86	2.33	43.03
	Shapley [17]	74.70 (-1.45↓)	92.18 (-0.69↓)	15.09	40.82	2.28	44.25
	FilterSketch-0.6 [24]	74.68 (-1.45↓)	92.17 (-0.69↓)	14.53	43.00	2.23	45.50
	FSE (Ours)	**74.81 (-1.34↓)**	**92.29 (-0.59↓)**	**14.03**	**45.10**	**2.21**	**46.60**
	AFPruner [35]	74.23 (-1.78↓)	91.89 (-1.07↓)	12.00	53.05	2.06	50.18
	FPGM [14]	74.13 (-2.02↓)	92.87 (-0.93↓)	-	-	1.92	53.50
	ABCPruner [34]	73.86 (-2.15↓)	91.69 (-1.27↓)	11.75	54.02	1.89	54.29
	Auto ℓ_0 Pruner-0.1 [18]	74.15 (-2.00↓)	91.70 (-1.18↓)	12.50	51.10	1.77	57.14
	FSE (Ours)	**74.43 (-1.77↓)**	**91.92 (-0.96↓)**	**12.92**	**49.43**	**1.76**	**57.53**
	HRank [15]	71.98 (-4.17↓)	91.01 (-1.86↓)	13.77	46.10	1.55	62.10
	FSE (Ours)	**73.12 (-3.03↓)**	**91.29 (-1.58↓)**	**12.16**	**52.41**	**1.55**	**62.45**

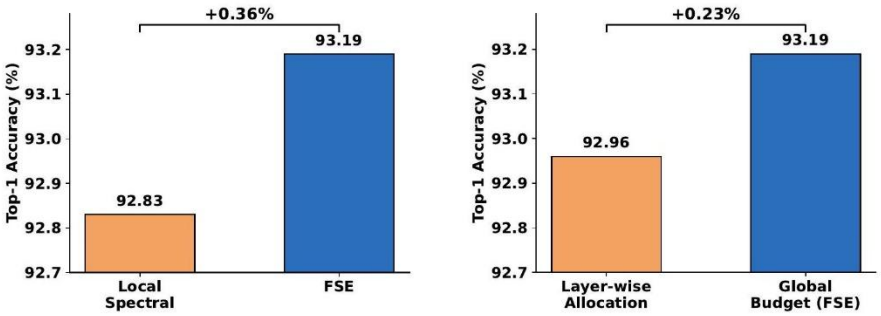
Compared with recent data-free pruning approaches, such as RASP [32] and CORING [31], FSE offers a different combination of structural information and pruning simplicity. RASP derives a data-independent amplitude-based saliency criterion, but its pruning framework involves additional regularization and model-selection steps followed by fine-tuning, whereas CORING uses higher-order SVD to identify redundant filters through within-layer similarity and adopts an iterative K-shot pruning strategy. In its main experiments, CORING uses $K = 15$, allowing fine-tuning between successive pruning shots, with a total of 100 calibration epochs allocated to the K-shot procedure [31]. In contrast, FSE determines the final pruning structure directly from the learned weights and performs fine-tuning only after the complete pruning decision. Therefore, the slightly higher accuracy achieved by RASP or CORING can be viewed in the context of their respective pruning protocols, which allow additional model adaptation during the pruning process. FSE instead targets a favorable accuracy-compression trade-off while maintaining a direct, lightweight and analytically defined pruning decision based solely on the learned weights.

4.3 Ablation Studies

We isolate the contribution of the two core components of FSE; the inter-layer spectral coupling cost and the global adaptive budget through controlled, single-variable ablations on ResNet-56/CIFAR-10. In each experiment, we change exactly one component and keep everything else identical: the SVD filter descriptors, the per-layer pruning caps, the skip-connection protection and the fine-tuning schedule. Crucially, every variant is evaluated at the same FLOP reduction (59.84%) rather than at the same nominal sparsity, because two criteria that prune the same number of filters generally remove different amounts of compute; matching FLOPs ensures that the accuracy gaps reported in Figs. 5a and 5b are not confounded by the operating point.

Effect of inter-layer spectral coupling. Replacing the coupling-based cost with a purely local perfilter spectral score, the descriptor norm $\|d_i\|_2$; i.e., each filter's own leading singular value, with the inter-layer coupling matrix discarded lowers Top-1 accuracy from 93.19% to 92.83% (-0.36 pp, Fig. 5a). Both variants use the same singular-value descriptors, so this gap cannot be attributed to the use of spectral information *per se*; it is specifically the modeling of how filter sub-spaces in adjacent layers couple that

yields the improvement. This directly supports our central claim that the importance of a filter is better measured by its role in the inter-layer spectral interaction than by its isolated magnitude.



(a) Effect of inter-layer spectral coupling (b) Effect of the global adaptive budget

Figure 5. Ablation study on ResNet-56/CIFAR-10 at a matched FLOP reduction. (a) Effect of the inter-layer spectral coupling. (b) Effect of the global adaptive budget.

Effect of the global adaptive budget. Replacing FSE's global budget with a uniform layer-wise allocation; the same pruning ratio in every eligible layer, with filters still removed cheapest-first within each layer, lowers Top-1 accuracy from 93.19% to 92.96% (-0.23 pp, Fig. 5b). Since the per-filter cost is identical in both runs, this gap measures the value of allocating the pruning budget globally: redistributing capacity across layers-pruning more aggressively where filters are spectrally redundant and sparing layers where they are not-outperforms a fixed uniform schedule at matched FLOPs.

Summary. The two components are complementary and both contribute positively, with the inter-layer coupling being the larger single contributor (-0.36 pp when removed) and the global budget providing a further, independent gain (-0.23 pp when removed). Together, they account for the accuracy advantage of FSE, confirming that neither the spectral descriptor alone nor a naive global ranking is sufficient; the gain stems from their combination.

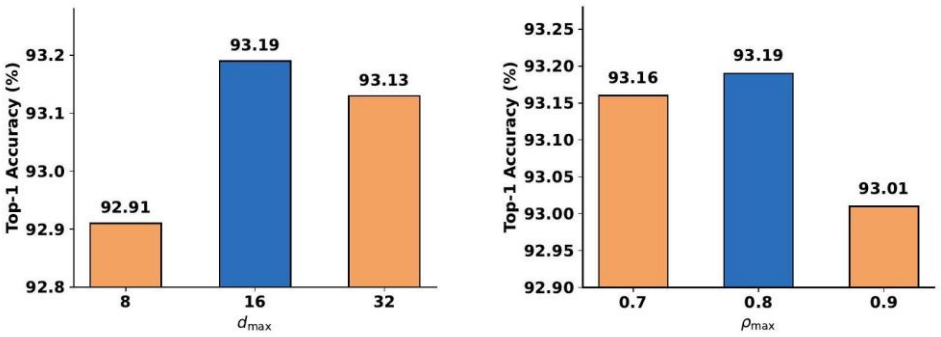
4.4 Pruning-decision Computational Overhead

The measured pruning overhead reported in Table 6, obtained on ResNet-56/CIFAR-10 at a matched FLOPs reduction of 59.84%, shows that the complete FSE pruning decision is obtained in only 0.940 s, excluding fine-tuning. The dominant cost is the construction of the SVD-based filter descriptors, which accounts for approximately 73.9% of the total time (694.73ms). This cost arises from computing the compact spectral representation of the learned filters and is incurred only once during the pruning decision. The subsequent construction of the coupling matrices requires 132.42ms (14.1%), while the SVD of the coupling matrices takes only 50.15ms (5.3%). The spectral-cost computation and global budget allocation are comparatively negligible, requiring 48.35ms (5.1%) and 14.83ms (1.6%), respectively. The peak CUDA memory consumption during importance scoring is only 16.0 MB, indicating that the coupling construction and its SVD do not introduce a substantial memory burden.

Table 6. FSE pruning overhead during importance scoring only. Fine-tuning time is excluded.

Pruning-decision component	Time (ms)
SVD filter descriptors (Eqs. 3-7)	694.73
Build coupling matrices (Eq. 10)	132.42
SVD of coupling matrices (Eq. 11)	50.15
Spectral costs (Eq. 12)	48.35
Global budget + ranking (Eqs. 13-15)	14.83
Total pruning-decision time	940.48 (0.940 s)
Peak CUDA memory during scoring	16.0 MB

Overall, these measurements show that although FSE involves additional spectral computations compared with simple criteria, such as weight norms, the complete pruning-decision stage remains lightweight, requiring less than one second and without any data-dependent forward passes. Importantly, this overhead is incurred only once before the subsequent fine-tuning stage and therefore does not constitute a recurring training cost.



(a) Sensitivity to the descriptor dimension $d_{\max}$ (b) Sensitivity to the per-layer pruning cap $\rho_{\max}$

Figure 6. Sensitivity analysis of FSE on ResNet-56/CIFAR-10 at a matched FLOP reduction. (a) Effect of the descriptor projection dimension $d_{\max}$. (b) Effect of the per-layer pruning cap $\rho_{\max}$.

4.5 Sensitivity to the Descriptor Dimension and Per-layer Cap

FSE exposes two design hyper-parameters: the descriptor projection dimension $d_{\max}$ in Eq. (8) and the per-layer pruning cap $\rho_{\max}$ in Eq. (14). We assess their influence by varying one hyper-parameter at a time at a matched FLOP reduction of 59.84%. The results are summarized in Figure 6.

Descriptor dimension $d_{\max}$. Top-1 accuracy varies by 0.28 pp across the tested range and reaches its highest value at $d_{\max} = 16$. Reducing the dimension to $d_{\max} = 8$ lowers accuracy to 92.91% (-0.28 pp), indicating that a more aggressive projection can discard information relevant to the spectral descriptors and their coupling. Increasing the dimension to $d_{\max} = 32$ does not improve the result, yielding 93.13% (-0.06 pp) relative to $d_{\max} = 16$. This suggests that increasing the descriptor dimension beyond 16 provides limited practical benefit under the tested pruning setting. Moreover, because $d_{\max}$ only affects the projected descriptor representation, its influence is limited to layers for which the original descriptor dimension exceeds the selected projection dimension. Overall, $d_{\max} = 16$ provides the best accuracy among the tested values without introducing an unnecessarily large descriptor representation.

Per-layer cap $\rho_{\max}$. The tested values produce a narrow accuracy range of 0.18 pp, with the highest accuracy obtained at $\rho_{\max} = 0.8$. Tightening the cap to $\rho_{\max} = 0.7$ results in only a 0.03 pp decrease (93.16%), indicating that a moderately stricter constraint has little effect on the final accuracy at the considered FLOP target. In contrast, increasing the cap to $\rho_{\max} = 0.9$ reduces accuracy to 93.01% (-0.18 pp). This result suggests that allowing substantially more aggressive pruning within individual layers can be less favorable, despite the same overall FLOP reduction, as it permits greater concentration of pruning in selected layers. The results therefore support $\rho_{\max} = 0.8$ as a balanced operating point between global allocation flexibility and protection against excessive layer-wise pruning.

Summary. Across the tested values, the maximum variation in Top-1 accuracy is 0.28 pp for $d_{\max}$ and 0.18 pp for $\rho_{\max}$. The relatively small variations indicate that FSE is reasonably robust to both hyper-parameters within the evaluated ranges. The selected values, $d_{\max} = 16$ and $\rho_{\max} = 0.8$, provide the best accuracy among the tested configurations while avoiding unnecessarily large descriptor dimensions or overly permissive layer-wise pruning.

5. CONCLUSION

This paper presented Filter Spectral Efficiency (FSE), a structured filter-pruning framework that estimates filter importance through the spectral coupling between consecutive convolutional layers. Unlike conventional criteria that evaluate filters independently or rely on activation statistics, FSE derives pruning decisions directly from the learned weights by modeling inter-layer information flow. The proposed framework further combines this relational spectral criterion with a global adaptive pruning budget that allocates sparsity automatically across the network while preserving structural consistency. Extensive experiments on five CNN architectures for CIFAR-10 and ResNet-50 on ImageNet demonstrated that FSE consistently achieves a favorable accuracy-compression trade-off, matching or outperforming representative structured pruning methods while requiring no calibration data or feature-map collection during the pruning stage. Ablation studies further confirmed that both the

inter-layer spectral criterion and the global adaptive budget contribute independently to the overall performance, validating the design choices of the proposed framework. These observations also suggest promising future directions for combining FSE with complementary compression techniques, such as tensor decomposition and knowledge distillation, as discussed in the positioning of this work. Future work will investigate extending the proposed spectral coupling formulation to transformer architectures and other modern vision models, incorporating hardware-aware objectives, such as latency and energy consumption, into the global budget and exploring more general forms of inter-layer dependency modeling for efficient neural network compression.

A. THEORETICAL PROPERTIES

This appendix collects the formal properties of the FSE construction. Unless stated otherwise, $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ denotes the thin singular value decomposition of a coupling matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, with $\mathbf{U} \in \mathbb{R}^{m\times r}$ satisfying $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_r$ and $\mathbf{\Sigma} = \mathrm{diag}(\sigma_1 \geq \cdots \geq \sigma_r \geq 0)$, where $r = \mathrm{rank}(\mathbf{M})$.

Lemma 1 (Information-preserving alignment). Let $\mathbf{c} \in \mathbb{R}^d$ and let $\tilde{\mathbf{c}} = [\mathbf{c}^\top, \mathbf{0}^\top]^\top \in \mathbb{R}^{d'}$ with $d' \geq d$ be its zero-padding. Then $\|\tilde{\mathbf{c}}\|_2 = \|\mathbf{c}\|_2$ and for any two descriptors padded to a common dimension the cosine similarity in Eq. (10) is independent of the padding target d'. Consequently the entries of $\mathbf{M}^{(l)}$ do not depend on the alignment dimension $d_\star$.

Proof. Appending zero coordinates leaves the sum of squares unchanged, so $\|\tilde{\mathbf{c}}\|_2^2 = \sum_j c_j^2 + 0 = \|\mathbf{c}\|_2^2$. For descriptors $\mathbf{a} \in \mathbb{R}^{d_a}$ and $\mathbf{b} \in \mathbb{R}^{d_b}$ padded to any $d' \geq \max(d_a, d_b)$, the inner product is $\langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}} \rangle = \sum_{j=1}^{\min(d_a,d_b)} a_j b_j$, since in every coordinate beyond $\min(d_a, d_b)$ at least one factor is zero; this value does not depend on d'. As both norms are likewise independent of d', so is the cosine $\langle \tilde{\mathbf{a}}, \tilde{\mathbf{b}} \rangle / (\|\tilde{\mathbf{a}}\|\|\tilde{\mathbf{b}}\|)$. Truncation, by contrast, would drop nonzero terms from the inner product and is therefore not informationpreserving.

Lemma 2 (Boundedness of the coupling spectrum).

$$\sigma_1 \leq \|\mathbf{M}\|_F \leq \sqrt{mn}, \sum_{k=1}^{r} \sigma_k = \|\mathbf{M}\|_* \leq \sqrt{r}\|\mathbf{M}\|_F \leq \sqrt{rmn}.$$

Proof. By definition $M_{ij} = \cos(\mathbf{d}_i, \mathbf{d}_j)$, hence $|M_{ij}| \leq 1$ by the Cauchy-Schwarz inequality and $\|\mathbf{M}\|_F^2 = \sum_{i,j} M_{ij}^2 \leq mn$. The spectral norm satisfies $\sigma_1 = \|\mathbf{M}\|_2 \leq \|\mathbf{M}\|_F$ and applying Cauchy-Schwarz to the vector of singular values gives $\sum_k \sigma_k \leq \sqrt{r}(\sum_k \sigma_k^2)^{1/2} = \sqrt{r}\|\mathbf{M}\|_F$. Combining these yields the stated bounds.

Proposition 1 (Spectral interpretation of the pruning cost). Let $\kappa_i = \sum_k U_{ik}^2 \sigma_k$ be the pruning cost defined in Eq. (12). Then

(i) $\boldsymbol{\kappa} = \mathrm{diag}((\mathbf{M}\mathbf{M}^\top)^{1/2})$, where $(\mathbf{M}\mathbf{M}^\top)^{1/2}$ is the unique positive semidefinite square root of $\mathbf{M}\mathbf{M}^\top$.

(ii) $\sum_i \kappa_i = \sum_k \sigma_k = \|\mathbf{M}\|_*$, that is, the pruning costs exactly decompose the nuclear norm of the coupling matrix.

(iii) For every filter, $0 \leq \kappa_i \leq \sigma_1$.

Proof. (i) Since $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$,

$$\mathbf{M}\mathbf{M}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top.$$

Define $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top$. The matrix $\mathbf{A}$ is symmetric and positive semidefinite and

$$\mathbf{A}^2 = \mathbf{U}\mathbf{\Sigma}(\mathbf{U}^\top\mathbf{U})\mathbf{\Sigma}\mathbf{U}^\top = \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top = \mathbf{M}\mathbf{M}^\top.$$

By uniqueness of the positive semidefinite square root, $\mathbf{A} = (\mathbf{M}\mathbf{M}^\top)^{1/2}$. Its diagonal entries satisfy $A_{ii} = \sum_k U_{ik}\sigma_k U_{ik} = \sum_k U_{ik}^2\sigma_k = \kappa_i$, which proves (i).

(ii) Taking the trace, $\sum_i \kappa_i = \mathrm{tr}(\mathbf{A}) = \mathrm{tr}(\mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top) = \mathrm{tr}(\mathbf{\Sigma}\mathbf{U}^\top\mathbf{U}) = \mathrm{tr}(\mathbf{\Sigma}) = \sum_k \sigma_k = \|\mathbf{M}\|_*$.

(iii) Let $w_k = U_{ik}^2 \geq 0$. Since $\mathbf{U}\mathbf{U}^\top$ is an orthogonal projector, $\sum_k w_k = (\mathbf{U}\mathbf{U}^\top)_{ii} \leq 1$. Hence $0 \leq \kappa_i = \sum_k w_k\sigma_k \leq \sigma_1 \sum_k w_k \leq \sigma_1$, which completes the proof.

Remark 1 (Sign invariance). The descriptor of Eq. (5) is defined only up to sign and Eq. (6) selects one representative. This choice does not affect the criterion. Let $\boldsymbol{S}_l = \mathrm{diag}\left(\varepsilon_0^{(l)}, \ldots, \varepsilon_{C_{\mathrm{out}}^{(l)}-1}^{(l)}\right)$ with $\varepsilon_i^{(l)} \in \{-1, +1\}$ be an arbitrary sign choice for layer l. Because the projection in $Eq.$ (8) is an uncentred SVD projection, replacing $\boldsymbol{D}^{(l)}$ by $\boldsymbol{S}_l\boldsymbol{D}^{(l)}$ leaves the retained principal subspace unchanged-indeed $\boldsymbol{D}^{(l)\top}\boldsymbol{S}_l$ has the same left singular vectors $\boldsymbol{U}_D$ and the same singular values as $\boldsymbol{D}^{(l)\top}$-and yields $\boldsymbol{S}_l\widetilde{\boldsymbol{D}}^{(l)}$. Zeropadding and row-wise ℓ_2 normalisation commute with $\boldsymbol{S}_l$ as well, so $\boldsymbol{M}^{(l)} \mapsto \boldsymbol{S}_l\boldsymbol{M}^{(l)}\boldsymbol{S}_{l+1}$.

Hence $\boldsymbol{M}\boldsymbol{M}^\top \mapsto \boldsymbol{S}_l\boldsymbol{M}\boldsymbol{M}^\top\boldsymbol{S}_l$ and, by uniqueness of the positive semidefinite square root, $(\boldsymbol{M}\boldsymbol{M}^\top)^{1/2} \mapsto \boldsymbol{S}_l(\boldsymbol{M}\boldsymbol{M}^\top)^{1/2}\boldsymbol{S}_l$. Conjugation by a diagonal sign matrix preserves diagonal entries, so Proposition 1(i) gives $\kappa_i^{(l)}$ unchanged for every filter i and every layer l. The pruning costs, the induced global ranking and the resulting pruned set are therefore independent of the sign convention.

Remark 2. Proposition 1 provides a direct spectral interpretation of the proposed pruning criterion. The quantity $(\mathbf{U}\mathbf{U}^\top)_{ii}$ represents the spectral leverage of filter i, while κ_i measures the contribution of that filter to the nuclear norm of the inter-layer coupling matrix. Consequently, filters with small values of κ_i contribute only weakly to the dominant inter-layer coupling modes and are therefore natural candidates for pruning.

Proposition 2 (Optimality of the global budget). Let $\mathcal{E} = \bigcup_l \mathcal{E}_l$ be the eligible filters partitioned by layer, with per-layer caps cap_l and let the costs $\{\kappa_i^{(l)}\}$ be arbitrary reals. The greedy rule that scans the filters in nondecreasing cost and admits a filter if and only if its layer has not yet reached cap_l, stopping after T admissions, returns a set $\mathcal{P}$ that (a) is feasible, i.e. $|\mathcal{P} \cap \mathcal{E}_l| \leq \mathrm{cap}_l$ for all l and (b) has minimum total cost among all feasible sets of cardinality $|\mathcal{P}|$.

Proof. The feasible pruning sets form a partition matroid whose independent sets are $\mathcal{I} = \{S \subseteq \mathcal{E} : |S \cap \mathcal{E}_l| \leq \mathrm{cap}_l, \forall l\}$.

The proposed global allocation is exactly the classical greedy algorithm on this partition matroid: filters are processed in non-decreasing pruning cost and selected whenever the corresponding layer capacity is not exceeded. By the fundamental optimality theorem for matroid greedy algorithms [41], the resulting set is a minimum-cost feasible subset for every attainable cardinality. With the stopping rule, the attained cardinality equals $\min(T, \sum_l \mathrm{cap}_l)$, so the selected pruning set $\mathcal{P}$ is both feasible and globally optimal under the prescribed layer-wise constraints.

Proposition 3 (Dimensional consistency of the structural surgery). Consider the supported architecture class in which all producer-consumer

dependencies are correctly identified, including sequential, residual and concatenation-based topologies. If, for every pruned layer l, the output channels indexed by $\mathcal{P}_l$ are removed from $\mathbf{W}^{(l)}$ and the same indices are removed from the input channels of every consumer of layer l (Eqs. (16)-(18)), then the pruned network admits a valid forward pass on any input of the original spatial resolution and every residual addition remains dimensionally defined.

Proof. The proof follows by induction over the computation graph in topological (data-flow) order. The first convolution is unchanged and therefore preserves the original input dimension. Assume that all predecessors of layer l satisfy the required channel consistency. Removing the output channels indexed by $\mathcal{P}_l$ from layer l and the corresponding input channels from every consumer preserves dimensional agreement along every producer-consumer edge; the accompanying batch-normalization parameters are indexed by the same retained set $\mathcal{K}_l$ and hence remain consistent. Residual additions remain valid because the protected branches in $\mathcal{S}$ preserve their original output dimensions, while concatenation operators receive tensors whose channel dimensions are consistently updated. Therefore every producer-consumer dependency remains dimensionally consistent, thereby guaranteeing that the structurally pruned network admits a valid forward pass.

REFERENCES

[1] E. Edozie et al., "Comprehensive Review of Recent Developments in Visual Object Detection Based on Deep Learning," Artificial Intelligence Review, vol. 58, no. 9, p. 277, 2025.

[2] M. H. Reza et al., "A Comprehensive Review of Convolutional Neural Networks: Foundations, Enhancements and Applications," Neural Computing and Applications, vol. 38, no. 4, p. 56, 2026.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 770-778, 2016.

[4] G. Huang, Z. Liu, L. van der Maaten and K. Q. Weinberger, "Densely Connected Convolutional Networks," Proc. of the IEEE Conf. on Comp. Vision and Patt. Recog. (CVPR), pp. 2261-2269, 2017.

[5] C. Szegedy et al., "Going Deeper with Convolutions," Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1-9, Boston, USA, 2015.

[6] H. Cheng, M. Zhang and J. Q. Shi, "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis and Recommendations," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 10558-10578, 2024.

[7] R. Schwartz, J. Dodge, N. A. Smith and O. Etzioni, "Green ai," Commun. ACM, vol. 63, no. 12, p. 54-63, [Online], Available: https://doi.org/10.1145/3381831, Nov. 2020.

[8] J. Yang et al., "A Survey on Neural Network Quantization," Proc. of the 2025 6th Int. Conf. on Computer Infor. and Big Data Appl. (CIBDA '25), Association for Comp. Machinery, p. 384-394, NY, USA, 2025.

[9] S. Zhang and R. Saab, "Theoretical Guarantees for Low-rank Compression of Deep Neural Networks," Applied and Computational Harmonic Analysis, vol. 82, p. 101837, 2026.

[10] T. Shehzadi et al., "Knowledge Distillation in Object Detection: A Survey from CNN to Transformer," Sensors (Basel, Switzerland), vol. 26, no. 1, p. 292, 2026.

[11] K. Zhu et al., "A Comprehensive Review of Network Pruning Based on Pruning Granularity and Pruning Time Perspectives," Neurocomputing, vol. 626, p. 129382, 2025.

[12] S. Han, J. Pool, J. Tran and W. J. Dally, "Learning both weights and connections for efficient neural networks," in Advances in Neural Information Processing Systems (NeurIPS), 2015, pp. 1135-1143.

[13] H. Li, A. Kadav, I. Durdanovic, H. Samet and H. P. Graf, "Pruning Filters for Efficient Convnets," Proc. of the Int. Conf. on Learning Representations (ICLR), arXiv: 1608.08710, 2017.

[14] Y. He, P. Liu, Z. Wang, Z. Hu and Y. Yang, "Filter Pruning *via* Geometric Median for Deep Convolutional Neural Networks Acceleration," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4340-4349, 2019.

[15] M. Lin etal., "HRank: Filter Pruning Using High-rank Feature Map," Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1529-1538, 2020.

[16] S. Lin et al., "Towards Optimal Structured CNN Pruning *via* Generative Adversarial Learning," Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2785-2794, 2019.

[17] T. Yuan, S. Yang, B. Liu and Y. Tang, "Explainable CNN Filter Pruning Based on Shapley Value Approximation in View of Game Theory," Expert Systems with Appl., vol. 300, p. 130232, 2025.

[18] E. Chen, H. Wang, Z. Shi and W. Zhang, "Channel Pruning for Convolutional Neural Networks Using l0-Norm Constraints," Neurocomputing, vol. 636, p. 129925, 2025.

[19] Ambuj, "Feature Representation Fidelity Preservation During Neural Network Pruning for Enhanced Compression Efficiency," Neurocomputing, vol. 634, p. 129901, 2025.

[20] A. Krizhevsky, "Learning Multiple Layers of Features from Tiny Images," Uni. of Toronto, Technical Report, [Online], Available: https://cave.cs.toronto.edu/kriz/learning-features-2009-TR.pdf, 2009.

[21] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," Proc. of the Int. Conf. on Learning Representations (ICLR), 2015.

[22] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," Int. Journal of Computer Vision, vol. 115, no. 3, pp. 211-252, 2015.

[23] Y. Sui et al., "Chip: Channel Independence-based Pruning for Compact Neural Networks," Advances in Neural Information Processing Systems, vol. 34, pp. 24604-24616, 2021.

[24] M. Lin, L. Cao, S. Li, Q. Ye, Y. Tian, J. Liu, Q. Tian and R. Ji, "Filter Sketch for Network Pruning," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 7091-7100, 2021.

[25] J.-H. Luo et al., "ThiNet: A Filter Level Pruning Method for Deep Neural Network Compression," Proc. of the IEEE Int. Conf. on Computer Vision, pp. 5058-5066, Venice, Italy, 2017.
[26] Y. He, X. Zhang and J. Sun, "Channel Pruning for Accelerating Very Deep Neural Networks," Proc. of the IEEE Int. Conf. on Computer Vision, pp. 1389-1397, Venice, Italy, 2017.
[27] M. Alwani et al., "DECORE: Deep Compression with Reinforcement Learning," Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, pp. 12349-12359, New Orleans, USA, 2022.
[28] Y. Zheng et al., "A Novel and Efficient Model Pruning Method for Deep Convolutional Neural Networks by Evaluating the Direct and Indirect Effects of Filters," Neurocomputing, vol. 569, p. 127124, 2024.
[29] Y. Lian, P. Peng, K. Jiang and W. Xu, "Cross-layer Importance Evaluation for Neural Network Pruning," Neural Networks, vol. 179, p. 106496, 2024.
[30] Y. He et al., "Soft Filter Pruning for Accelerating Deep Convolutional Neural Networks," Proc. of the 27th Int. Joint Conf. on Artificial Intelligence (IJCAI), pp. 2234-2240, 2018.
[31] V. T. Pham, Y. Zniyed and T. P. Nguyen, "Efficient Tensor Decomposition-based Filter Pruning," Neural Networks, vol. 178, p. 106393, 2024.
[32] C. Zhen, W. Zhang, J. Mo, M. Ji, H. Zhou and J. Zhu, "Rasp: Regularization-based Amplitude Saliency Pruning," Neural Networks, vol. 168, pp. 1-13, 2023.
[33] Z. Liu et al., "Learning Efficient Convolutional Networks through Network Slimming," Proc. of the 2017 IEEE Int. Conf. on Computer Vision (ICCV), pp. 2755-2763, Venice, Italy, 2017.
[34] M. Lin, R. Ji, Y. Zhang, B. Zhang, Y. Wu and Y. Tian, "Channel Pruning *via* Automatic Structure Search," Proc. of the 29th Int. Joint Conf. on Artificial Intelligence (IJCAI'20), pp. 673-679, 2021.
[35] Y. Xue, W. Yao, S. Peng and S. Yao, "Automatic Filter Pruning Algorithm for Image Classification," Applied Intelligence, vol. 54, no. 1, pp. 216-230, 2024.
[36] V. T. Pham et al., "Enhanced Network Compression through Tensor Decompositions and Pruning," IEEE Trans. on Neural Networks and Learning Systems, vol. 36, no. 3, pp. 4358-4370, 2025.
[37] V. T. Pham et al., "Coupled Tensor Decomposition for Compact Network Representation," IEEE Trans. on Neural Networks and Learning Systems, vol. 37, no. 2, pp. 617-631, 2026.
[38] W. Lin et al., "STPP: Efficient and Progressive Structured Pruning *via* Enhanced Sparsification Paradigm," IEEE Trans. on Pattern Analysis and Machine Intelligence, Early Publishing, pp. 1-17, 2026.
[39] G. Fang et al., "DepGraph: Towards Any Structural Pruning," Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 16091-16 101, Vancouver, Canada, 2023.
[40] Z. Huang and N. Wang, "Data-driven Sparse Structure Selection for Deep Neural Networks," Proc. of the 15th Europ. Conf. on Comp. Vision (ECCV 2018), Part 16, pp. 317-334, Munich, Germany, 2018.
[41] A. Schrijver, Combinatorial Optimization: Polyhedra and Efficiency, E-ISSN: 2197-6783, Springer, 2003.

ملخص البحث:

يُعدّ تقليم المرشّحات المهيكل نهجاً فعّالاً لتقليل التّكلفة الحسابية للشّبكات العصبية العميقة مع الحفاظ على إمكانية النّشر الفعّال على العتاد القياسي. ومع ذلك، فإنّ معظم طرق التّقليم الحالية تقيم المرشّحات بشكلٍ مستقلّ، متجاهلةً الاقتران الطّيفي بين الطّبقات المتتالية، في حين تعتمد الأساليب عالية الأداء على إحصائيات التّنشيط أو بيانات المعايرة.

نقترح في هذه الورقة "الكفاءة الطّيفية للمرشّحات" عبر إطار عملٍ للتّقليم المهيكل لا يعتمد على البيانات، حيث يقدّر أهمّية المرشّح من خلال الاقتران الطّيفي بين الطّبقات المتجاورة. ويعمل الإطار المقترح على إنشاء واصِفاتٍ للمرشّحات تعتمد على القيم المفردة، وبناء مصفوفات الاقتران بين الطّبقات، واشتقاق معيارٍ للتّقليم الطّيفي، وتوزيع التّناثُر عبر ميزانية تكييفية شاملة.

وقد أظهرت التّجارب المكثّفة التي أجريت على خمس بنياتٍ للشّبكات العصبية التّلافيفية أنّ الإطار المقترح يحقّق باستمرارٍ توازناً مُمتازاً بين الدّقّة ومعدّل الضّغط، متفوّقاً بذلك على الطّرق النّموذجية للتّقليم المهيكل عند مستوياتٍ مماثلةٍ من تقليل المعْلمات والعمليات الحسابية. وبينت النّتائج أنّ الرّبْط الطّيفي بين الطّبقات يوفّر أساساً فعّالاً وعملياً لتّقليم الشّبكات المهيكل دون الحاجة إلى بياناتٍ.

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات (JJCIT) مجلة علمية عالمية متخصصة محكمة تنشر الأوراق البحثية الأصيلة عالية المستوى في جميع الجوانب والتقنيات المتعلقة بمجالات تكنولوجيا وهندسة الحاسوب والاتصالات وتكنولوجيا المعلومات.

تحتضن وتنشر جامعة الأميرة سمية للتكنولوجيا (PSUT) المجلة الأردنية للحاسوب وتكنولوجيا المعلومات، وهي تصدر بدعم من صندوق دعم البحث العلمي في الأردن. وللباحثين الحق في قراءة كامل نصوص الأوراق البحثية المنشورة في المجلة وطباعتها وتوزيعها والبحث عنها وتنزيلها وتصويرها والوصول إليها. وتسمح المجلة بالنسخ من الأوراق المنشورة، لكن مع الإشارة إلى المصدر.

الأهداف والمجال

تهدف المجلة الأردنية للحاسوب وتكنولوجيا المعلومات (JJCIT) إلى نشر آخر التطورات في شكل أوراق بحثية أصيلة وبحوث مراجعة في جميع المجالات المتعلقة بالاتصالات وهندسة الحاسوب وتكنولوجيا المعلومات وجعلها متاحة للباحثين في شتى أرجاء العالم. وتركز المجلة على موضوعات تشمل على سبيل المثال لا الحصر: هندسة الحاسوب وشبكات الاتصالات وعلوم الحاسوب ونظم المعلومات وتكنولوجيا المعلومات وتطبيقاتها.

الفهرسة

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات مفهرسة في كل من:

فريق دعم هيئة التحرير

المحرر اللغوي	ادخال البيانات وسكرتير هيئة التحرير
حيدر المومني	إياد الكوز

جميع الأوراق البحثية في هذا العدد مُتاحة للوصول المفتوح، وموزعة تحت أحكام وشروط ترخيص [Creative Commons Attribution] (http://creativecommons.org/licenses/by/4.0/)

عنوان المجلة

الموقع الإلكتروني: www.jjcit.org

البريد الإلكتروني: jjcit@psut.edu.jo

العنوان: جامعة الاميرة سمية للتكنولوجيا، شارع خليل السكاكت، الجبيهة، عمان، الأردن.

صندوق بريد: 1438 عمان 11941 الأردن

هاتف: 962-6-5359949+

فاكس: 962-6-7295534+

جامعة الأميرة سميّة للتكنولوجيا
Princess Sumaya University for Technology

صندوق دعم البحث العلمي والابتكار
Scientific Research and Innovation Support Fund

المجلة الأردنية للحاسوب وتكنولوجيا المعلومات

أيلول ٢٠٢٦ المجلد ١٢ العدد ٣
ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

JJCIT

الصفحات	عنوان البحث
٢٦٦ - ٢٨٢	نحو دمْجٍ موثوقٍ لواجهات برمجة تطبيقات إنترنت الأشياء والويب: منهجية تركيب البيانات المرتبطة مع مراعاة التّاريخ محمد أمين بالفضل، ميمون مالكي، و عبد الرحمن خياط
٢٨٣ - ٢٩٦	رؤى تحليلية حول متوسّط معدّل خطأ الكتلة في أنظمة الهوائيات المائعة في ظلّ قيود الحُزم القصيرة باو تشي لي، و تين فو تران
٢٩٧ - ٣١٦	تطوّرات الكشف عن اعتلال الشّبكية السُّكري المدعوم بالذّكاء الاصطناعي: دراسة مسحية منهجية للطّرق والتّحديات والتّوجهات المستقبلية ياسر أ. علي، أحمد ز. إمام، محمد ح. زيان، سمير الموجي، و أحمد س. طُلبة
٣١٧ - ٣٣٣	مُحاذاة النّماذج الأوّلية ثنائية المستوى عبر آلية الانتباه المتقاطع للتّجزئة الدّلالية في الاستشعار عن بُعدٍ باستخدام عيّنات قليلة مصطفى العوادي، و منصور فاتح
٣٣٤ - ٣٥٩	التّنبُّؤ بأخطاء البرمجيات باستخدام نموذجٍ هجين هرمي مبتكر مهراسا م. جويباري، علي رضا تاجاري، منصور فاتح، و إسماعيل تهانيان
٣٦٠ - ٣٨٧	أداء الشّبكة لنظام التّرحيل ثنائي الاتّجاه مع ترميز «النّافورة» في شبكات الاستشعار اللّاسلكية في ظلّ وجود تداخل في القناة المشتركة ثي-أنه نغو، و كوك-هاو ترونغ
٣٨٨ - ٤١١	تصنيف آفات الثّدي: التّعلُّم بالنّقل القائم على الرّقع والخالي من تسرُّب البيانات، مع تقييمٍ باستخدام عددٍ من مجموعات البيانات هيتش كاغ، دينيش كومار، و أنجالي ديوان
٤١٢ - ٤٢٨	الكفاءة الطّيفية للمرشّحات: تقليم الشّبكات المهيكل عبر الاقتران الطّيفي بين الطّبقات سمية لاروي، حميدة سيبا، كريمة عمروش، و ياسين آيت علي يحيى

www.jjcit.org
jjcit@psut.edu.jo

مجلة علمية عالمية متخصصة تصدر
بدعم من صندوق دعم البحث العلمي والابتكار