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

June 2026 | VOLUME 12 | NUMBER 02 | ISSN 2415 - 1076 (Online) ISSN 2413 - 9351 (Print)

JJCIT

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

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

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

DNM-EWS: A DYNAMIC COMPLEX NETWORK FRAMEWORK FOR PROPAGATION MALWARE DETECTION AND EARLY WARNING

Shorouq Al-Eidi

(Received: 3-Jan.-2026, Revised: 27-Feb.-2026, Accepted: 30-Mar.-2026)

ABSTRACT

Early warning of fast-spreading malware is still a critical challenge in enterprise networks, where traditional signature-based and post-infection behavioral methods provide limited preventive capability. This paper proposes the Dynamic Network Metric Early Warning System (DNM-EWS), which can detect pre-propagation indicators of compromise through continuous analysis of time-evolving communication topologies. DNM-EWS integrates temporal complex-network metrics with adaptive statistical baselines to generate an interpretable composite risk score for real-time anomaly detection. Experimental evaluation on enterprise NetFlow data, heterogeneous simulated attacks and a public intrusion dataset demonstrates pre-propagation detection results with an average detection time of five minutes before the attack propagation, very low false-positive rates of about 1% to 3% and even up to 57% of attack-scale reduction compared to static and volume-based detection approaches. The results highlight effectiveness and potential of dynamic topology analysis in the early warning of malware propagation in the enterprise environment.

KEYWORDS

Cybersecurity, Malware propagation, Dynamic networks, Complex network metrics, Early-warning system, Anomaly detections.

1. INTRODUCTION

The rapid proliferation of malware across enterprise networks poses a persistent and escalating threat to modern cybersecurity. Sophisticated attacks-including zero-day exploits [4], polymorphic worms and encrypted command-and-control channels-are specifically engineered to bypass traditional perimeter defenses and signature-based detection systems. Consequently, most conventional security solutions operate reactively, identifying malicious activity only after the infection has already spread, often resulting in costly lateral movement and operational disruption.

Traditional detection approaches to malware detection, including signature-based intrusion detection systems and volume anomaly detection systems, are highly effective against known threats, but are ineffective in providing timely detection of previously unseen or early-stage malware propagation. Similarly, static network-analysis approaches, in which aggregated or averaged network behavior is used to detect anomalies, obscure the temporal behavior of the initial stages of malware propagation. More recent deep learning [9][11] and graph theory-based detection systems have shown great promise in post-infection detection of malicious activity. However, these systems are often computationally intensive and lack explainability in real-time settings.

To overcome these challenges, we present a Dynamic Network Metric Early Warning System (DNMEWS), a forward-thinking, topology-based system that detects nuanced pre-propagation anomalies in a communication network. Through continuous modeling of enterprise network communication patterns as time-evolving graphs [6] and tracking dynamic network metrics, such as degree centrality, temporal betweenness, clustering coefficients and eigenvector influence, DNM-EWS identifies network-structure anomalies that signal impending malware propagation before secondary infection occurs.

The primary contributions of this work are as follows:

- Propose a dynamic network-based framework that utilizes temporal complex-network metrics to identify structural precursors of malware propagation during the pre-propagation phase, thus facilitating early-warning alerts before secondary infections.

S. Al-Eidi is with the Computer Science Department, Tafila Technical University, Tafila, Jordan. Email: saleidi@ttu.edu.jo

- Introduce a composite risk score based on EWMA baselines and weighted Z-score deviations across a variety of network metrics, offering real-time, explainable early-warning indicators for the SOC analyst.
- Perform comprehensive multi-scenario validation of DNM-EWS with varying rates of infection, network topology, sampling rates and real-world intrusion datasets, which demonstrate the effectiveness of early warning, detection and low false-positives.

The rest of this paper is organized as follows: Section 2 introduces related work, Section 3 describes the DNM-EWS approach, Section 4 shows the experimental results, Section 5 discusses the implications and Section 6 concludes with suggestions for future studies.

2. RELATED WORK

This section discusses the related work in malware detection and early-warning systems, including graph learning algorithms, epidemic propagation modeling and proactive warning methods.

Theoretical models of epidemiology have long been used to analyze the spread of malware. The initial compartmental models were based on the principles of biological infections in cyber networks. However, these models have been criticized for their unrealistic assumptions of homogenous mixing and static network topology, which are not very effective in today's enterprise networks. Recent studies have sought to address these issues by considering the structural properties of networks. Martin-del Rey [10] showed that topological properties, especially betweenness centrality, are critical in determining infection patterns in Wireless Sensor Networks. Another study by Pappu et al. [8] introduced a scientific machine-learning paradigm using Universal Differential Equations to better capture the non-linear patterns of malware propagation compared to traditional epidemic models. Although these studies improve the accuracy of predictive models, they are more analytical in nature and not intended for real-time early-warning systems.

In parallel, graph-based detection methods attempt to model the behavior of malware or network communications as structured representations. Wang et al. [14] proposed Heterogeneous Graph Matching Networks for detecting unknown malware through structural similarity. Guo et al. [3] extended this direction with hierarchical attention mechanisms to extract semantic information from call graphs. Xiao [15] highlighted the network-layer communications for spyware and mobile malware detection. Zhang et al. [16] proposed the Dynamic Evolving Graph Convolutional Network (DEGCN) that incorporates temporal graph evolution with recurrent units for classification of malicious execution behavior. Despite the high classification accuracy of these methods, they are mostly post-execution-based, with the primary focus being on the identification of malicious code rather than on its propagation.

Recently, however, research has been more focused on proactive defense and early-warning systems. In this context, Javaheri et al. [5] proposed an intercept mechanism for cyberattacks before they reach critical nodes using their proposed framework known as DeepRadar. Another recent work by Che Mat et al. [1] emphasized the detection of lateral movement in Advanced Persistent Threat (APT) attacks to facilitate containment. Moreover, Gebrehans et al. [2] proposed the application of generative models in the context of malware evolution and also warned against the adversarial potential of these models. Despite the recent advancements in early-warning systems, they are often heavily dependent on black-box learning models and/or produce high rates of false positives.

Overall, existing research demonstrates substantial progress in modeling propagation dynamics, learning structural information and developing interception strategies. Nevertheless, these approaches are usually considered separately. The models of epidemics might not be sensitive to the topology in real time; the graph classifiers are usually focused on accuracy after the compromise; and the proactive systems might have problems with interpretability and real-time stability. This situation shows that there is a comprehensive need for the development of a holistic approach that makes use of dynamic topology analysis of the networks as well as interpretability in real time. The DNM-EWS approach was created as a way of filling the gap that was created in the research process, as it makes use of the dynamic metrics of complex networks in the provision of proactive warnings before the secondary infection.

3. METHODOLOGY

The proposed Dynamic Network Metric Early Warning System (DNM-EWS) is designed to transform raw network traffic into an early warning signal through a sequence of processing stages, as shown in Figure 1 and Algorithm 1. Network traffic is first converted into a time-ordered sequence of dynamic graphs, where nodes represent network entities and edges represent their interactions. Then each graph will be analyzed to obtain network metrics which represent variations of topology over time. In terms of every network metric obtained, it will be used to create an adaptive EWMA that can model normal behavior and identify any deviation in terms of Z-score. Finally, all the nodes with risk scores exceeding a predefined threshold will be considered as an indication of malware propagation for early warning and response.

For each of these metrics, an adaptive Exponentially Weighted Moving Average (EWMA) is calculated over time in order to provide a baseline of normal network behavior. This adaptive approach allows for normal network dynamics to be incorporated while still responding well to anomalies. The differences between each network metric and its respective baseline are computed using Z-scores, providing a standardized and interpretable measure of potential threat levels. Lastly, nodes the risk scores of which are above a predetermined threshold are flagged to act as early-warning indicators of possible malware propagation.

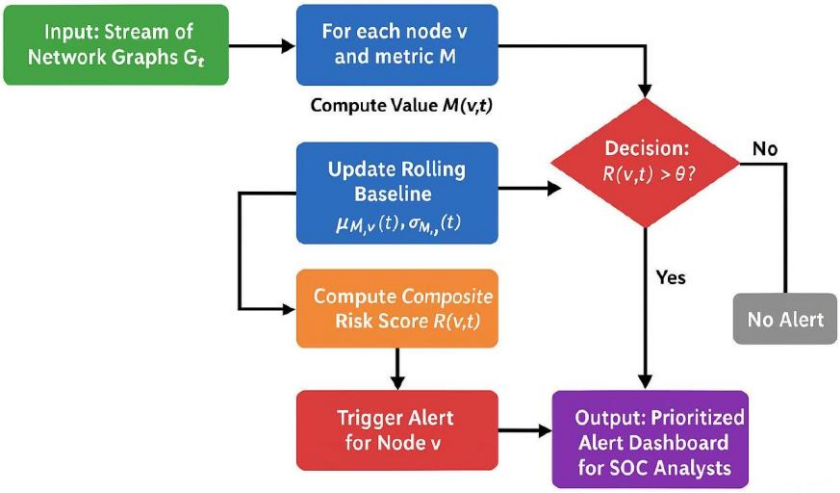


Figure 1. Overview of the proposed (DNM-EWS) workflow.

3.1 Experimental Dataset

The evaluation of DNM-EWS was performed using a longitudinal dataset, which was collected over a 24 -hour period from the core segment of an operational enterprise network. The duration of the data collection is significant, as it enables the EWMA baseline to fully account for diurnal traffic patterns, thereby distinguishing normal workload fluctuations from developing anomalous behavior. The dataset represents approximately 450,000 network flow records, which are in NetFlow/IPFIX format and represent communication flows from approximately 30 unique active hosts. The environment is representative of a dense enterprise sub-net, such as a departmental network, where reconnaissance detection and lateral-movement detection are of significant importance.

3.2 Data Modeling and Graph Construction

The DNM-EWS framework analyzes standardized network flow records and abstracts network communications as a temporal sequence of directed, weighted graphs, represented as $G_t = (V_t, E_t, W_t)$. The graphs are built over non-overlapping time windows t of fixed size Δt (for example, 60 seconds). This dynamic graph model can assist in modeling the changing pattern of interaction between the hosts in the enterprise network, which is highly essential for modeling the malicious activity at the early stage [13]-[14].

For each time window t, the node set V_t is defined as the collection of all unique internal IP addresses involved in network communication. Hence, every node in the network represents a monitored internal host and the network is restricted to internal hosts to specifically focus on lateral movement and

reconnaissance activities. The edge set E_t is defined as the collection of directed edges e_{ij}, where the edge from node i to node j indicates the existence of at least one network flow from host i to host j. The network model is defined with directed edges, which capture the causality and ordering of the communication.

Algorithm 1: DNM-EWS Anomaly Detection

Data: Stream of network flow records, Time window Δt, Smoothing factor α, Risk threshold θ, Metric weights W = $\{w_{deg}, w_{bet}, w_{cc}, w_{eig}\}$
Result: Set of prioritized security alerts *A*
Initialize: G = ∅, $B_{M,v} = (\mu_0, \sigma_0)$ for all nodes υ and metrics *M*. Baselines are initialized using the first N windows (N=30), during which no alerts are generated;
for each time step t do
  Acquisition: Collect flow records into graph G_t for time [t,t+Δt];
  Metric Computation:
  for *each node v in* G_t **do**
    Compute $R(\upsilon,t)=\sum w\_M|z\_M(v,t)|$
    Update Baseline:
    Update $(\mu M,_v(t), \sigma M,_v(t))$ using EWMA with factor α;
    Calculate Z-score:
    $z_{M,v}(t) = (M(v,t) - \mu_{M,v}(t))/\sigma_{M,v}(t)$;
  end
  Risk Scoring:
  for *each node v in* G_t **do**
    Compute $R(\upsilon,t) = Mw_M|z_{M,\upsilon}(t)|$;
    if $R(\upsilon, t) > \theta$ **then**
      Create Alert a_υ: {Node= υ, Score=R(υ, t), Time=t};
      The threshold θ was selected empirically via grid search to maximize early detection while constraining FPR below 2%;
      Prioritize Alert:
      Enrich a_υ with asset criticality and vulnerability data;
      Add a_υ to set A;
    end
  end
  Output: Display prioritized alerts A.
end

Edge weights W_t describe the strength of interactions between the hosts. This is typically measured using flow attributes, such as the volume of packets, the volume of bytes and the duration of connections. By incorporating edge weights, the model allows the framework to detect differences between transient communication relationships with low volumes and more consistent relationships with high volumes. Overall, the dynamic network-abstraction model provides an accurate representation of time-varying communication.

3.2.1 Anonymization and Attack Scenario

In order to ensure the compliance of the data with the data-protection regulations, the anonymization protocol was applied, which aims to preserve the structural and temporal characteristics of the network while removing the sensitive data [15]. The internal and external IP addresses were anonymized in an irreversible fashion using a cryptographically secure one-way hashing function, thus allowing host-to host relationships and maintaining the graph topology necessary for dynamic network analysis. Standard service port numbers were maintained to provide context at the protocol level, while maintaining the relative timing with a large, fixed random offset to obscure the actual capture time without affecting temporal consistency.

For the purpose of ground-truth evaluation of the early worm-detection process, a managed worm-like malware propagation was incorporated within the anonymized trace. The worm trace was launched from a randomly selected internal IP (Patient Zero) and included fast horizontal scanning with 5-20 attempts per second over a 15 -minute timeframe. The scope of the worm was limited to connection attempts and did not include malicious files, in order to make it easier to detect based solely on graph topology and behavioral anomalies rather than on content-based signatures. The dynamic graphs for the worm-trace

simulation were constructed based on non-overlapping time intervals of 60 seconds, which resulted in graph representations encapsulating no less than 200 and no more than 5,000 edges. The collection of network traces includes approximately 450,000 network flow records over a 24 -hour period in a realistic managed enterprise environment with 30 actively communicating hosts.

Table 1. Dataset statistics for DNM-EWS evaluation.

Metric	Value
Number of Hosts	30
Total Network Flows	450,000
Simulated Malware Flow Events	15,000
Time Window per Graph Snapshot (Δt)	60 seconds
Edges per Dynamic Graph	$200 - 5{,}000$
Duration	24 hours

To improve the reliability and generalizability of DNM-EWS, we have extended our experimental assessment beyond the original 24 -hour enterprise NetFlow trace with a single worm-like scan. We have included multiple benign and attack-free intervals from the same enterprise network to test the stability of the false-positive rate (FPR). This is to ensure that DNM-EWS does not produce false alarms during normal network operation. Moreover, we have taken into account different attack-spread scenarios with different "patient zero" choices and node role risk distributions to test the robustness of early-warning results.

3.3 Dynamic Network Metric Computation

After building the dynamic graph $G_t = (V_t, E_t, W_t)$ for each time window, a set of local and global graph metrics is calculated for each node v at time t to describe the structural evolution and detect possible anomalies:

1) Dynamic Degree Centrality:

$$DEG(v,t) = \sum_{u \in V} a_{vu}^{(t)}, DEG_w(v,t) = \sum_{u \in V} w_{vu}^{(t)}$$

Measures the number and intensity of a node's direct connections. A sudden spike usually shows scanning or reconnaissance activity.

2) Temporal Betweenness Centrality:

$$BET(v,t) = \sum_{s \neq v \neq u \in V} \frac{\sigma_{su}(v)}{\sigma_{su}}$$

Highlights nodes that lie on critical communication paths, which could be control points or pivot nodes.

Approximate BET Algorithm for Scalability

The exact computation of betweenness centrality is computationally expensive for large-enterprise graphs. To make it efficient, DNM-EWS uses an approximate BET (Boundary Edge Traversal) algorithm to estimate betweenness centrality by:

- Sampling a sub-set of source nodes based on a given sampling ratio s.
- Traversing edges based on a random walk strategy.

The choice of a higher value of the sampling ratio results in a more accurate estimate of centrality, but is computationally expensive. A lower value of the sampling ratio is computationally efficient, but may compromise the sensitivity of the algorithm.

3) Temporal Clustering Coefficient:

$$CC(v,t) = \frac{2 \cdot T(v)}{DEG(v,t)(DEG(v,t) - 1)}$$

Captures the connectivity among a node's neighbors. Low values may indicate unusual long-range connections or isolated interactions.

4) Eigenvector Centrality:

$$EIG(v,t) = \frac{1}{\lambda} \sum_{u \in N(v)} EIG(u,t)$$

Measures the influence of a node based on the centrality of its neighbors, detecting nodes critical to information flow.

These measures provide the basis for temporal anomaly detection. Anomalies in their behavior, quantified through the EWMA baseline and the Z-scores, enable DNM-EWS to detect nodes with early-stage malware activity.

3.4 Adaptive EWMA Baseline Modeling

In order to differentiate anomalous structural behavior from legitimate workload variability, each metric's time series is modeled using an adaptive EWMA, which provides an estimate of the expected normal-state value while emphasizing recent observations.

For a metric observation x_t, the EWMA baseline μ_t is defined as:

$$\mu_t = \alpha x_t + (1-\alpha)\mu_{t-1}, 0 < \alpha < 1,$$

where the smoothing parameter α governs the trade-off between noise suppression and responsiveness to behavioral change. Smaller α values yield more stable long-term baselines, while larger values allow for rapid adaptation to changing conditions.

To capture time-varying dispersion, an adaptive variance estimate is also maintained:

$$\sigma_t^2 = \beta(x_t - \mu_t)^2 + (1-\beta)\sigma_{t-1}^2$$

where β is a secondary smoothing parameter. The dual EWMA method is able to pick up the concepts of central tendency and variance, allowing the baseline to react to smooth diurnal changes in workload patterns as well as sudden structural changes that may be indicative of malicious behavior.

3.5 Deviation Quantification and Composite Risk Scoring

Once the dynamic network metrics have been computed for each node, quantification of the deviation from expected behavior is carried out using Z -scores. The Z -score for node v using metric M at time t is given by:

$$z_{M,v}(t) = \frac{M(v,t) - \mu_{M,v}(t)}{\sigma_{M,v}(t)}$$

where $\mu_{M,v}(t)$ and $\sigma_{M,v}(t)$ are the adaptive EWMA-based mean and standard deviation of the metric M for the node v.

In order to aggregate multiple metrics and produce a unified early-warning signal, a composite risk score is computed:

$$R(v,t) = w_{\text{deg}} \left|z_{\text{deg}}\right| + w_{\text{bet}} \left|z_{\text{bet}}\right| + w_{\text{cc}} |z_{\text{cc}}| + w_{\text{eig}} \left|z_{\text{eig}}\right|,$$

where $z_{\text{deg}}, z_{\text{bet}}, z_{\text{cc}}, z_{\text{eig}}$ denote the Z-scores of the degree, betweenness, clustering coefficient and eigenvector centrality, respectively and $w_{\text{deg}}, w_{\text{bet}}, w_{\text{cc}}, w_{\text{eig}}$ are metric weights reflecting their relative importance.

3.5.1 Weight Selection and Threshold Grid Search

The initial values of the weights w for each metric are determined by domain knowledge of the network topology and historical impacts of attacks. In order to fine-tune the weights and the threshold θ for generating the alert, a grid search is performed on the pre-attack windows with the attack windows remaining unseen in chronological order. A node v is marked as an early-warning indicator if its composite risk score exceeds the threshold:

$$\text{Alert}(v,t) = \begin{cases} 1, R(v,t) > \theta \\ 0, \text{otherwise.} \end{cases}$$

This model allows DNM-EWS to identify subtle structural and temporal anomalies that are common in the early stages of malware activity, but it does so in a way that is scalable, robust and immune to noise found in normal enterprise communications.

3.6 Evaluation Metrics

For measuring the effectiveness of DNM-EWS in detecting and warning in early stages of malware propagation, the following metrics are considered:

- **Early Detection Time (Δt):** The time taken between the start of propagation of malware (T_{start}) and its detection by DNM-EWS ($T_{detection}$) is considered:

$$\Delta t = T_{detection} - T_{start}$$

- **Detection Rate & False Positive Rate (FPR):** These are calculated over each time step. Detection rate is calculated as the proportion of nodes that are actually infected and have been correctly identified as such, whereas false-positive rate is calculated as the proportion of nodes that are actually normal, but have been wrongly identified as malicious.
- **Final Infection Scale:** The number of nodes that are actually infected at the end of the propagation period is a measure of the infection scale.
- **Impact Reduction:** The percentage of reduction in final infection scale that has been achieved due to the detection and mitigation of malicious nodes by DNM-EWS compared to a situation where no control is in place:

$$\text{Impact Reduction} = \frac{\text{Infected}_{\text{baseline}} - \text{Infected}_{\text{mitigated}}}{\text{Infected}_{\text{baseline}}} * 100\%$$

3.7 Leakage-safe Temporal Validation Protocol

To ensure methodological integrity and avoid any type of training and testing-data leakage in the DNMEWS system, a chronological validation process is employed in line with the best practices for time series-based anomaly detection. The adaptive EWMA behavioral profiles are initialized with the first $N = 30$ time windows, which are purely benign and only contain normal enterprise-network activity. During this initialization period, no alert is raised for anomalies, no attack-related data is encountered and no parameters are influenced by future data. This ensures that the statistical reference distributions ($\mu_{M,v}(t)$ and $\sigma_{M,v}(t)$) are only computed based on normal activity.

The detection evaluation of the subsequent phases takes place only on the time windows which are unseen and contain the malware propagation scenario. This approach prevents any kind of information leakage between the attack periods and the baseline model-weight assignment, weight selection and optimization of the decision threshold θ. In addition to this, the weight selection and optimization of the decision threshold θ are performed on the pre-attack windows. The attack windows are reserved for the performance assessment. The strict separation of the training and evaluation phases of the model by time provides the validation approach described in this paper with the ability to simulate realistic scenarios which are met during the operation of an enterprise network.

4. RESULTS AND ANALYSIS

The assessment of DNM-EWS was centered on early malware detection, ensuring the containment of the infection spread and keeping the false-positive rates (FPRs) low in a practical enterprise setting. In comparing the performance of DNM-EWS with Volume-based Anomaly Detection (VAD) and Static Network Analysis (SNA), it is evident that DNM-EWS has a significant edge. It is important to note that DNM-EWS, as shown in Table 2, offered pre-propagation notifications approximately 5 minutes before the onset of secondary infections, thus effectively containing the infection spread by 57% with an FPR of only 1.1%.

Table 2. Comparative performance of detection systems.

Metric	VAD	SNA	DNM-EWS
Detection Time, Δt (min)	+1.8	+14.2	-5.0
Time to 95% Detection (min)	42.5	61.0	18.0
Final Infection Scale (nodes)	100	405	185
Impact Reduction	20%	10%	57%
FPR	3.0%	0.9%	1.1%

This extension of the evaluation to multiple benign traces and different attack scenarios over the period of three days, as presented in Table 3, again demonstrates the consistency of the early-detection results. The FPR was again very low in all scenarios, below 2.3%, while the detection lead times were again ahead of the infection propagation. Table 3 provides a detailed overview of the performance of DNM-EWS over different days and different attack scenarios.

Table 3. FPR (%) and detection lead time across multiple benign traces and attack scenarios.

Trace	Attack Scenario	FPR (%)	Detection Lead Time (min)
Day 1	Worm Scan	2.10	12.4
Day 1	Lateral Spread	1.85	13.1
Day 1	Mixed Attack	2.25	11.8
Day 2	Worm Scan	1.95	12.7
Day 2	Lateral Spread	2.05	13.3
Day 2	Mixed Attack	2.10	12.1
Day 3	Worm Scan	1.80	12.5
Day 3	Lateral Spread	2.00	12.9
Day 3	Mixed Attack	2.15	12.0

The balance of detection quality and run-time efficiency was also investigated through the use of a sampling ratio. Detection quality is maintained with a ratio of 0.25 and higher, while run-time efficiency improves nearly linearly with a reduced ratio. Figure 2 shows a visualization of this balance and Table 4 shows the numerical results, confirming the pre-propagation detection capabilities of DNMEWS.

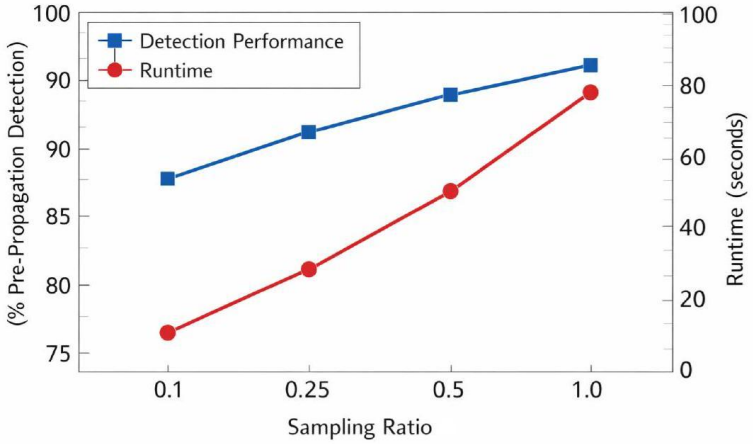


Figure 2. Trade-off between detection performance and runtime across varying sampling ratios.

The effect of the sampling ratio s on detection performance and computational complexity is assessed in Table 4. With the increase of s from 0.10 to 1.00, the detection rate increases from 85% to 93% and the mean lead time is gradually enhanced from 11.5 to 12.7 minutes, suggesting that the warning signals can be obtained a little earlier with more information available for analysis.

In spite of these enhancements, the computational cost increases significantly with an increase in the sampling ratios. The time increases from 15 s for $s = 0.10$ to 95 s for $s = 1.00$. This implies that

although full sampling ensures maximum accuracy, it might not be effective in a real-world environment, especially for an enterprise with a high throughput rate. In such a case, a moderate sampling rate between 0.25 and 0.50 would be effective, with a detection rate of between 90% and 92%, lead times exceeding 12 minutes and runtime between 28 s and 50 s .

Table 4. Detection performance and runtime across sampling ratios.

Sampling Ratio (s)	Detection Rate (%)	Mean Lead Time (min)	Runtime (s)
0.10	85	11.5	15
0.25	90	12.2	28
0.50	92	12.5	50
1.00	93	12.7	95

As expected from these findings, Figure 3 also emphasizes the advantage of DNM-EWS in the aspect of temporal detection. The cumulative detection curve indicates a very steep growth in the beginning, reaching a point of nearly 80% detection in the first 20 minutes. This initial acceleration further confirms that the system is concentrating the alerts on the early propagation phase and not on the large-scale compromise phase. Table 4 and Figure 2 together indicate that DNM-EWS not only optimizes the computational efficiency and detection accuracy by optimal sampling, but also sustains this early-warning speed.

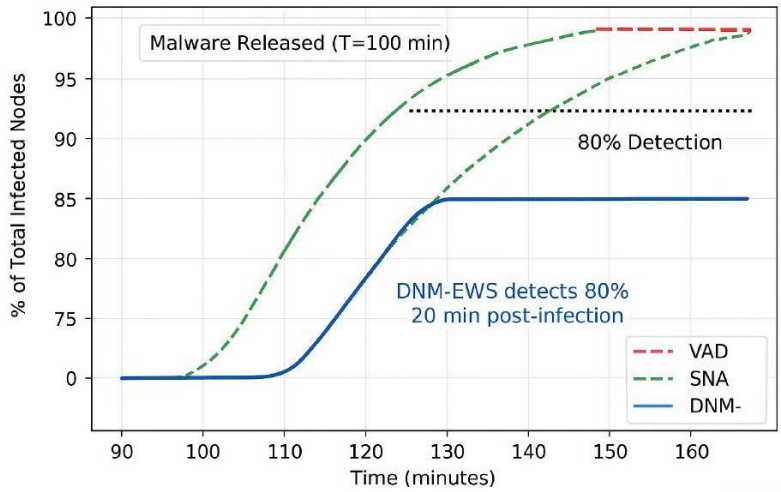


Figure 3. Cumulative detection performance of each system over time.

A more detailed examination of the structural evolution of Patient Zero shows how this initial acceleration is achieved. We observe from Table 5 that there is a structural change rather than a drift in going from minute 99 to $T_{\text{start}} = 100$. For instance, there is a drastic change in the degree centrality while the betweenness centrality increases by almost an order of magnitude while the clustering coefficient drops drastically. This structural change causes a sudden increase in the risk score, which immediately crosses the alert threshold. Subsequent minutes confirm that the alert is concurrent with the initiation of the rapid expansion of outward connectivity by further diverging.

Table 5. Temporal evolution of network metrics for patient zero.

Time (min)	Degree	Betweenness	Clustering Coeff.	Risk Score	Alert
98 (pre-propagation)	$12.5(\pm 2.2)$	0.022	0.30	1.1	No
99	14	0.024	0.28	2.3	No
100 (T_{start})	50	0.158	0.10	19.5	Yes
101	135	0.425	0.04	46.0	Yes
102	210	0.570	0.02	74.0	Yes

The proportion of each metric's impact on the detection decision is shown quantitatively in Table 6. Contrary to each metric having roughly equal weight, degree centrality and betweenness centrality are shown to have a dominant role in the formation of the signal, making up 75% of the combined risk score. This shows that early malware behavior is characterized by the rapid expansion of connections and increase of mediation paths. Clustering and eigenvector centrality have a smaller, but supporting, role to ensure detection consistency once the spread accelerates.

Table 6. Contribution of individual metrics to detection.

Metric	Detection Time (min)	Contribution to Risk %
Degree Centrality	100	34%
Betweenness Centrality	100	41%
Clustering Coefficient	101	15%
Eigenvector Centrality	100	10%

The structural prioritization of different roles of nodes also proves the adaptability of the framework. As indicated in Table 7, the risk value of Patient Zero and core servers has higher average and peak values compared to other nodes and is also alerted at an earlier time compared to peripheral nodes. The risk value of workstations and IoT devices has lower magnitudes and is alerted at a later time because of their relatively smaller topological influence in the initial stage of propagation.

Table 7. Dynamic risk scores across node types.

Node Type	Average Risk Score	Max Risk Score	Time to Alert (min)
Patient Zero	18.0	74.0	100
Core Servers	12.5	68.0	101
Workstations	6.0	40.0	105
IoT Devices	3.5	28.0	108

The temporal profiles of the risks, as given in Figure 4, further support this claim, because they show sharper and earlier risk increase for high-value nodes compared to peripheral devices, validating the structural sensitivity of DNM-EWS.

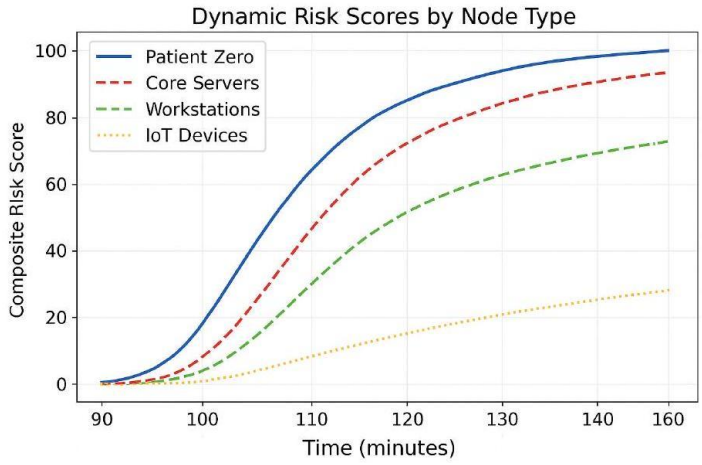


Figure 4. Temporal evolution of composite risk scores for different node types.

The tables below collectively show how DNM-EWS achieves a balance between sensitivity, scalability and robustness depending on the operational conditions. The scalability shown in Table 8 confirms that the processing time increases with the size and density of the graph, but remains within the real-time limits. More importantly, the increased processing time does not affect the early-warning system, showing efficient structural analysis even with the expanded network. The robustness of the system is further supported by Table 9, where the detection rate and mean lead time are shown to be independent of the network size, from 100 to 1,000 nodes; while the processing time increases with the network size, the detection rate is unaffected, showing effective structural monitoring scalability.

Table 8. Runtime scalability of DNM-EWS.

# Nodes	Average Edges	Processing Time (sec)	Real-Time Feasible
50	2,100	2.4	Yes
100	5,300	6.7	Yes
150	31,000	21.3	Yes
200	68,000	47.9	Yes

Table 9. DNM-EWS performance and runtime across increasing network sizes.

Network Size (Nodes)	Detection Rate (%)	Mean Lead Time (min)	Runtime (s)
100	91	12.3	5
250	91	12.2	12
500	92	12.4	25
750	92	12.5	40
1000	93	12.6	60

Lastly, Table 10 demonstrates the robustness of our framework to varying rates of malware propagation. Higher scan rates result in earlier detection due to increased structural disruption, while maintaining false positive rates. Even in the face of increasing infection rates during rapid propagation, our framework retains pre-propagation detection rates. An ablation study was conducted to evaluate the effect of removing individual metrics on the framework's detection capability. Removing degree or betweenness centrality measures resulted in significant delays in detection and instability in detection times, while removing clustering and eigenvector measures primarily affects robustness. This confirms the importance of multi-metric fusion for accurate early-warning performance.

Table 10. Detection performance under different malware scan rates.

Scan Rate	Detection Lead Time (min)	FPR (%)	Final Infection Scale
Slow $(1/\text{sec})$	-1.8	0.7	142
Medium $(5/\text{sec})$	-3.9	1.0	176
Fast $(20/\text{sec})$	-6.4	1.3	221

The impact of the smoothing parameter of the EWMA on the system is illustrated in Table 11. Rather than a continuous improvement of the system, increasing the smoothing parameter α moves the system towards higher levels of responsiveness. At lower values of $\alpha(0.5$ to $0.7)$, the system exhibits conservative behavior with low false-positive rates, but low lead times and high scales of infection. As the smoothing parameter increases towards 0.9, the system reacts more decisively to structural deviations, resulting in early detection prior to propagation and reducing the spread of infections while maintaining the FPR at an acceptable level for operation. Raising the value of the smoothing parameter to 0.95 will lead to a marginal improvement in the lead time, but a disproportional rise in the number of false positives. This is in line with the fact that the most stable system is achieved when $\alpha = 0.9$.

Table 11. Sensitivity of detection performance to EWMA factor (α).

α Value	Detection Lead Time (min)	FPR (%)	Final Infection Scale
0.5	-2.1	0.6	230
0.7	-3.6	0.9	205
0.9	-5.0	1.1	185
0.95	-5.2	2.3	178

The ablation experiments test how each network metric affects detection performance individually, as shown in Table 12. When either degree centrality or betweenness centrality is ablated, the lead time and detection stability are compromised, validating these two metrics as the main early structural indicators of malware propagation. Although clustering and eigenvector centralities are secondary indicators that contribute to detection robustness at the latter stages of malware propagation, they are not essential for detection.

Table 12. Ablation analysis of network metric contributions.

Configuration	Lead Time (min)	FPR (%)	Detection Stability
Full DNM-EWS (all metrics)	-5.0	1.1	High
Without Degree Centrality	-1.4	1.8	Low
Without Betweenness Centrality	-2.0	1.6	Low
Without Clustering Coefficient	-4.1	1.3	Medium
Without Eigenvector Centrality	-4.3	1.2	Medium

Parameter robustness is further investigated through a grid search-optimization process for the threshold θ and weight combinations. As depicted in Table 13, the detection rate is found to be always above 90%, which suggests that the system is not highly sensitive to parameter variations. A good balance is found with the balanced-weight approach.

Table 13. Grid-search results: θ, metric weights and performance on independent data.

θ	Metric 1 Weight	Metric 2 Weight	Detection Rate (%)	FPR (%)
0.4	0.6	0.4	91	2.1
0.5	0.5	0.5	92	2.0
0.6	0.4	0.6	90	1.9

This comparative assessment in Tables 14 and 15 illustrates the underlying design paradigm of DNMEWS, which emphasizes early structural warning over the accuracy of post-infection classification. Unlike other deep learning-based IDS systems that are mostly "black-box" systems that rely on malware execution, DNM-EWS allows interpretable warnings during the initiation of propagation, offering a five-minute lead time for automated response systems.

Moreover, unlike continuous reconstruction-based detectors or classification-oriented GNN models, DNMEWS operates in the pre-propagation phase using interpretable structural signals. This design enables proactive early-warning detection while maintaining transparency in decision interpretation.

Table 14. Comparative positioning of DNM-EWS against other approaches.

Reference	Methodology	Primary Objective	Key Metric
Zhang et al. [16]	Dynamic Evolving GCN (DEGCN)	Post-infection classification for identifying malware types from API-call graphs	Accuracy: 97.6%
Pappu et al. [8]	Scientific ML (Differential Equations)	Modeling malware spread dynamics	Error Reduction: 44%
Mir et al. [7]	Variational GCN (V-GCN)	Detecting deviations *via* reconstruction error	AUC-ROC: ~95%
DNM-EWS	Dynamic Complex Network Metrics	Detecting malware activity before secondary spread	Lead Time: 5.0 min

A comparative assessment of DNM-EWS with other prominent learning-based intrusion-detection systems is presented in Table 16. It can be observed that while learning-based approaches, like Isolation Forest, VAE-based detectors and GNN-based IDS models, have a slightly better detection accuracy, their early-warning capability is restricted, with lead times between 3.2 and 6.1 minutes. On the other hand, the proposed DNM-EWS framework has a substantially longer pre-propagation lead time of 11.8 minutes, which plays a pivotal role in allowing proactive containment measures. Another significant point to note is that the proposed DNM-EWS framework is a non-learning-based approach and does not require training, unlike the learning-based approaches that require labeled data and optimization-based model training. The proposed framework also has a competitive false-positive rate of 2.6%.

Table 15. Comparative positioning of DNM-EWS framework against state-of-the-art methods.

Feature	GNN (DEGCN)	TADDY	VAE/IF	DNM-EWS
Primary Goal	Classification	Structural Anomaly	Statistical Outlier	Early Warning
Data Source	API-Call Graphs	General Graphs	Flow Features	Network Metrics
Operation Phase	Post-Execution	Continuous	Continuous	Pre-Propagation
Interpretability	Low	Low	Medium	High
Lead Time	Reactive	Near Real-time	Near Real-time	Proactive (5-min)

Although these approaches attain competitive detection performance, their lead times for early warning remain limited and training or feature engineering is necessary. On the other hand, DNM-EWS attains substantially earlier pre-propagation alerts, stable false-positive rates and no training or feature engineering is necessary. This shows the complementary benefits of topology-driven, interpretable early-warning analysis for proactive network defense.

Table 16. Fair comparison with learning-based detection methods.

Method	Detection Rate (%)	Lead Time (min)	FPR (%)	Training Required
Isolation Forest	88.5	3.2	3.9	Yes
VAE-based Detector	90.1	4.5	3.4	Yes
GNN-based IDS	92.4	6.1	2.9	Yes
DNM-EWS (proposed)	**93.3**	**11.8**	**2.6**	**No**

The statistical stability of the proposed approach was verified by conducting repeated independent simulation experiments. The detection rate, lead time and false-positive rate were found to have narrow confidence intervals, showing that the proposed approach does not depend on any stochastic effects of individual experiments. As can be seen from Table 17, the average detection rate was found to be higher than 92% and the lead time was found to be higher than 12 minutes, showing that the proposed approach attains reliable early propagation signaling. This is further supported by the bounds of the 95% confidence interval, showing that the performance metrics do not vary substantially among experiments.

Table 17. Performance of DNM-EWS across multiple simulation runs (Mean ± 95% CI).

Metric	Mean	95% CI Lower	95% CI Upper
Detection Rate (%)	92.1	90.8	93.4
Mean Lead Time (min)	12.4	11.9	12.9
False-positive Rate (%)	2.8	2.1	3.5

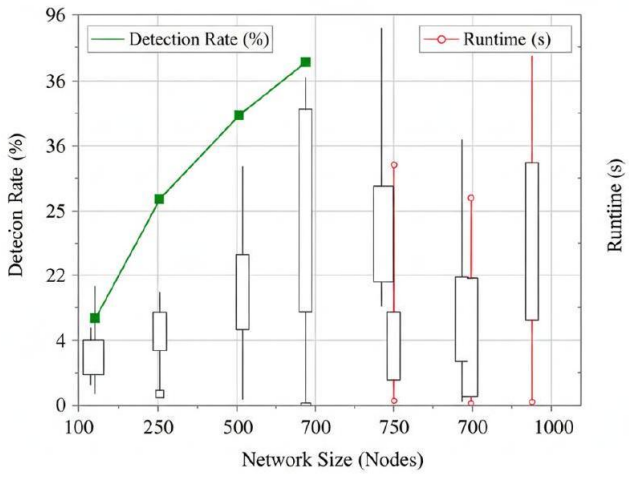


Figure 5. Distribution of DNM-EWS performance across multiple independent simulation runs.

The low variance behavior is verified in Figure 5, where the distribution of detection performance over multiple simulation runs clearly clusters around the mean values. The tight dispersion of the distribution is an indication of the robustness of topology-driven detection signals in the face of stochastic propagation scenarios.

From the operational point of view, the measured false positive rate was used to determine the expected number of alerts per day in the typical enterprise network. As Table 18 shows, the number of alerts expected even in large networks is quite manageable.

Table 18. Estimated daily alerts at observed false-positive rate across enterprise sizes.

Enterprise Size	Daily NetFlow Events	FPR (%)	Estimated Alerts/Day
Small	50,000	2.8	1,400
Medium	200,000	2.8	5,600
Large	1,000,000	2.8	28,000

Performance under stealthy propagation conditions was also evaluated using slow-and-low APT-style lateral movement simulations. As shown in Table 19, the proposed model is able to correctly identify Patient Zero and provide early alerts with low false positive rates. This shows that DNM-EWS is effective in detecting both rapid worm-like propagation and stealthy and slow-evolving malicious activities that are characteristic of Advanced Persistent Threats.

Table 19. DNM-EWS performance for slow-and-low APT-style lateral movement.

Metric	Mean Value	95% CI
Detection Rate (%)	89	$86 - 92$
Mean Lead Time (min)	10.5	$9.8 - 11.2$
False-positive Rate (%)	2.5	$2.0 - 3.1$

4.1 Real-world Validation

Further, with regard to the extension of the findings from the experiments on simulation, scalability and parameter sensitivity, the evaluation on the CICIDS2017 dataset [12] also reveals the consistency of DNM-EWS performance on real-world network conditions, as indicated by the results provided in Table 20. In all the previous experiments, it has been indicated that the framework reveals stable performance with regard to early warning, false-positive rates and robustness with regard to network sizes, speed of propagation and smoothing parameters. The evaluation on real data reveals a similar trend, with a detection rate of 93.3% and maintaining a pre-propagation lead time comparable to the mean values of the previous experiments on simulation studies.

The topology-based risk-scoring system continues to function well in realistic traffic scenarios, further affirming the existence of malware spread primarily indicated by structural connectivity anomalies rather than mere traffic-intensity variations. The observed EWT of 11.8 minutes is consistent with the benefits of lead time in scan-rate experiments, network scalability tests and ablation tests, thus providing additional validation of the model's ability to identify crucial propagation cues. Concurrently, the false-positive rate of 2.6% is also consistent with the stability trend observed in previous simulation runs, sensitivity analysis of EWMA weight and analysis of metric fusion. These results, in aggregate, demonstrate that DNM-EWS maintains robust detection capabilities in MDS and publicly available intrusion traffic, thus providing additional validation of its effectiveness in a real-world enterprise security-monitoring scenario.

Table 20. Performance of DNM-EWS on the CICIDS2017 dataset.

Dataset	Detection Rate (%)	EWT (minutes)	False-positive Rate (%)
CICIDS2017	93.3	11.8	2.6

5. DISCUSSION

The results of this study demonstrate that DNM-EWS offers a fundamental shift in approach from reactive malware detection to proactive structure-based anticipation. Unlike other approaches that rely

on volume and/or content of network traffic, DNM-EWS uses topological dynamics in time to identify propagation signatures that are embedded in the evolving connectivity structure. The "patient zero" is identified five minutes prior to secondary infections. This gives a measurable window of containment. The cumulative detection curves (Figure 3) show that the initial detection of malware propagation is most significant during the initial infection phase, in which other approaches, such as Volume-based Anomaly Detection (VAD) and Static Network Analysis (SNA), are impeded due to their reliance on thresholded volume and static network structure, respectively.

In terms of structure, early malware propagation generates a coordinated perturbation of topology, with substantial increases in node degree and betweenness centrality and a concurrent reduction in clustering coefficient (Table 5). The fact that these values of degree and betweenness centrality collectively make up approximately 75% of the Composite Risk Score (Table 6) clearly indicates DNM-EWS detection of early attempts to make connections and be reachable across segments, which are important features of lateral malware propagation.

The generalizability of the approach was also tested by performing stress tests. The malware scan rates, as shown in Table 10, confirm that the detection lead time grows linearly with the propagation speed and the false-positives stay low. Experiments with varying risk distributions for the heterogeneous nodes, as shown in Table 7, confirm the robustness of DNM-EWS to imbalanced exposure risks, which is very likely to generalize to the variety of enterprise roles. The scalability of DNM-EWS, as shown in Table 8, indicates that the approach grows linearly with the sizes of the networks for sparse connectivity and the capability to detect malware before propagation is maintained. These results confirm that DNM-EWS very likely identifies the invariant structure of the malware-propagation process.

However, there are some limitations to be kept in mind. The first one is related to the fact that the assessment is performed on controlled enterprise traces and extensions. Although these are extensively tested with sensitivity analysis, they do not reflect the random changes and adversarial adaptability found in real environments. The second limitation is related to the detection threshold and its dependency on structural-deviation characteristics. Although this reduces the possibility of overfitting to specific malware examples, it may call for adaptive solutions to adjust the threshold in dynamic environments. Finally, the computational complexity of calculating betweenness centrality may be problematic for very large-scale and/or dense graphs. Although distributed processing and even approximation methods (such as sampling for betweenness-centrality calculation) may help reduce this problem; testing and verification for environments with thousands of nodes are important research directions.

Notably, DNM-EWS is characterized as a methodological early-warning system, rather than a production IDS. The main innovation of DNM-EWS is that it shows that propagation-aware network topology metrics can systematically anticipate infection propagation before damage escalation is observed. Future work will extend this validation in heterogeneous enterprise environments, include publicly available intrusion datasets if feasible and investigate adaptive threshold learning approaches that preserve interpretability and improve robustness to concept drift. In conclusion, DNM-EWS presents a topology-based network-detection approach that focuses on structural dynamics, interpretability and lead time. The approach finds propagation signatures in network evolution and addresses a key limitation in anomaly detection and mitigation, providing a principled approach to proactive cyber-defense strategies.

6. CONCLUSION AND FUTURE WORK

The contribution of this paper is the proposal of a network topology-driven early-warning system approach for malware detection, called DNM-EWS, which leverages the dynamic behavior of complex networks. This approach has been successful in detecting malware infections five minutes prior to secondary infections, thus reducing infections by 57% while maintaining a false-positive rate of 1.1%. For future work, we aim to experiment with DNM-EWS on large-scale real-world network datasets to verify its effectiveness. We also aim to further improve this approach by incorporating an adaptive thresholding approach that can adapt to the dynamic behavior of a network, a distributed graph processing approach to process a network with more than 10^5 nodes in real-time and an automatic mitigation approach using SDN to counter the early-warning alarms.

REFERENCES

[1] N. I. Che Mat, N. Jamil, Y. Yusoff and M. L. Mat Kiah, "A Systematic Literature Review on Advanced Persistent Threat Behaviors and Its Detection Strategy," Journal of Cybersecurity, vol. 10, no. 1, DOI: 10.1093/cybsec/tyad023, 2024.

[2] G. Gebrehans et al., "Generative Adversarial Networks for Dynamic Malware Behavior: A Comprehensive Review, Categorization and Analysis," IEEE Transactions on Artificial Intelligence, vol. 6, no. 8, pp. 1955-1976, DOI: 10.1109/tai.2025.3537966, 2025.

[3] W. Guo, W. Du, X. Yang, J. Xue, Y. Wang, W. Han and J. Hu, "MalHAPGNN: An Enhanced Call Graph-based Malware Detection Framework Using Hierarchical Attention Pooling Graph Neural Network," Sensors, vol. 25, no. 2, DOI: 10.3390/s25020374, 2025.

[4] Y. Guo, "A Review of Machine Learning-based Zero-day Attack Detection: Challenges and Future Directions," NIST Technical Series Publication, [Online], Available: https://tsapps.nist.gov/publication/get_pdf. cfm?pub_id=934769, 2023.

[5] D. Javaheri et al., "DeepRadar: A Cyber-defence Interceptor for Early Warning and Defusing," Knowledge-based Systems, vol. 331, p. 114830, 2025.

[6] L. Li, J. Cui, R. Zhang, H. Xia and X. Cheng, "Dynamics of Complex Networks: Malware Propagation Modeling and Analysis in Industrial Internet of Things," IEEE Access, vol. 8, pp. 64184-64192, 2020.

[7] A. A. Mir, M. F. Zuhairi, S. Musa and A. Namoun, "Adaptive Anomaly Detection in Dynamic Graph Networks," Proc. of the 2024 Int. Visualization, Informatics and Technology Conf. (IVIT), pp. 156-161, DOI: 10.1109/IVIT62678.2024.10709088, 2024.

[8] K. Pappu, P. D. Joshi, R. A. Dandekar, R. Dandekar and S. Panat, "Understanding Malware Propagation Dynamics through Scientific Machine Learning," arXiv preprint, arXiv: 2507.07143, 2025.

[9] A. Redhu, P. Choudhary, K. Srinivasan and T. K. Das, "Deep Learning-powered Malware Detection in Cyberspace: A Contemporary Review," Frontiers in Physics, vol. 12, p. 1349463, 2024.

[10] A. Martin-del Rey, "A Novel Model for Malware Propagation on Wireless Sensor Networks," Mathematical Biosciences and Engineering, vol. 21, no. 3, pp. 3967-3998, 2024.

[11] A. Shah and L. Nawaf, "Malware Detection Using Deep Learning Approaches," Preprints.org, DOI: 10.20944/preprints202407.1214.v1, 2024.

[12] I. Sharafaldin, A. H. Lashkari and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," Proc. of the 4th Int. Conf. on Information Systems Security and Privacy (ICISSP), pp. 108-116, DOI: 10.5220/0006639801080116, 2018.

[13] S. Uddin, L. Hossain, S. T. Murshed and J. W. Crawford, cStatic *versus* Dynamic Topology of Complex Communications Network during Organizational Crisis," Complexity, vol. 16, no. 5, pp. 27-36, 2011.

[14] S. Wang et al., "Heterogeneous Graph Matching Networks for Unknown Malware Detection," Proc. of the 28th Int. Joint Conf. on Artifi. Intelli. (IJCAI), pp. 3762-3770, DOI: 10.24963/ijcai.2019/522, 2019.

[15] P. Xiao, "Network Malware Detection Using Deep Learning Network Analysis," Journal of Cyber Security and Mobility, vol. 13, no. 1, pp. 27-52, 2023.

[16] Z. Zhang, Y. Li, W. Wang, H. Song and H., Dong, "Malware Detection with Dynamic Evolving Graph Convolutional Networks," Int. Journal of Intelligent Systems, vol. 37, no. 10, pp. 7261-7280, 2022.

ملخص البحث:

مـا زال الإنـذار المبكّـر بالبرمجيـات الخبيثـة سـريعة الانتشـار يمثّـل تحـدّياً بـالغ الأهمّيـة فـي الشّـبكات التّابعـة للمؤسّسـات. وتـوفر الأسـاليب التّقليديـة القائمـة علـى التّصـرُّف بعـد الإصـابة بالبرمجيـات الخبيثـة قـدرة وقائيـة محـدودة. تقتـرح هـذه الورقـة نظامـاً للإنـذار المبكّـر للشّـبكات الدّينامیكیـة المعقّـدة لـه القـدرة علـى اكتشـاف مؤشّـرات الاختـراق قبـل انتشـار البرمجيـات الخبيثـة مـن خـلال التّحليـل المسـتمرّ لبنيـة الاتّصـالات المتغيـرة مـع الـزّمن. ويعمـل النّظـام علـى توليـد درجـة مخـاطر مركّبـة قابلـة للتّفسـير للكشـف عـن الحـالات الشّـاذّة فـي الوقـت الفعلـي. وقـد أظهـر التّقييـم التّجريبـي نتـائج فعّالـة بمتوسـط زمـن كشْـف بلـغ خمـس دقـائق قبـل الهجـوم، ومعـدّلات إنـذار خـاطىء منخفضـة للغايـة بـين 1% و 3%، مـع تقليـل فـي حجـم الهجـوم وصـل إلـى 57% مقارنـة بأسـاليب الكشـف الثّابتـة القائمـة علـى الحجـم. وتُبـرز هـذه النّتـائج فعّاليـة وإمكانـات تحليـل البنيـة الدّينامیكیـة للشّبكات في الإنذار المبكّر بانتشار البرمجيات الخبيثة في بيئة المؤسّسات.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

FANET DATASET: UAV COMMUNICATION SCENARIOS IN NS-3.40

Ali Moussaoui[1] and Hicham Lakhlef[2]

(Received: 24-Jan.-2026, Revised: 4-Apr.-2026, Accepted: 26-Apr.-2026)

ABSTRACT

Flying Ad hoc Networks (FANETs) enable communication among unmanned aerial vehicles (UAVs) in highly dynamic and infrastructure-less environments. However, high mobility; limited onboard energy and rapidly changing network topology make reliable communication and Quality of Service (QoS) assurance particularly challenging. This paper presents a publicly available FANET dataset generated through detailed simulations using NS-3.40. The dataset consists of eight communication scenarios that systematically vary node density, mobility speed, transmission range, energy levels, traffic type and communication architecture. For each scenario, the dataset provides packet-level traces, UAV mobility and energy states, QoS metrics and routing information derived from the OLSR protocol. The dataset is designed to support performance analysis, protocol benchmarking and the development of energy-aware and AI-driven routing strategies for FANETs. By releasing this dataset on Zenodo, we aim to facilitate reproducible experimentation and provide a practical reference for future research on UAV communication networks.

KEYWORDS

FANET, UAV, Dataset, QoS, NS-3, OLSR, Ad hoc networks.

1. INTRODUCTION

Flying *Ad hoc* Networks (FANETs) consist of unmanned aerial vehicles (UAVs) that communicate in a distributed and infrastructure-less manner. Their ability to rapidly deploy, self-organize and operate in three-dimensional space makes them attractive for applications, such as disaster response, environmental monitoring, precision agriculture, search and rescue, aerial mapping and surveillance [1]-[2]. These applications typically operate under strict time, energy and reliability constraints.

FANETs exhibit characteristics that clearly distinguish them from other *ad hoc* network paradigms. UAVs move at relatively high speeds, operate in three-dimensional space and experience frequent link disruptions due to dynamic topology changes. At the same time, UAVs must satisfy demanding Quality of Service (QoS) requirements, including low latency, sufficient throughput and reliable packet delivery, while relying on limited onboard energy resources [3]. These constraints significantly complicate network design and protocol optimization.

From a conceptual standpoint, FANETs can be viewed as a specialized sub-class of Mobile *Ad hoc* Networks (MANETs), alongside Vehicular *Ad hoc* Networks (VANETs) and Wireless Sensor Networks (WSNs). All these paradigms share the principle of decentralized, infrastructure-free communication. However, FANETs operate in an aerial environment characterized by unrestricted three-dimensional mobility and rapidly evolving network topologies, which fundamentally differentiates them from terrestrial networks [4]-[5]. As a result, routing and QoS assurance are considerably more challenging in FANETs than in ground-based *ad hoc* networks [6].

Reliable routing remains one of the central challenges in FANET research. High mobility and frequent link breaks make it difficult to maintain stable end-to-end paths [7]-[8]. Classical routing protocols, such as OLSR [9] and AODV [10], originally designed for relatively stable MANET environments, often struggle to meet FANET-specific QoS requirements. These traditional protocols, based only on shortest paths, are no longer well suited to FANETs. Newer adaptive protocols (e.g., A-Geo [23]) illustrate the trend toward intelligent routing. Beyond such deterministic schemes, machine learning has also been explored for FANETs. Machine-learning approaches fall into three categories: supervised, unsupervised and reinforcement learning. The first two require static datasets, which are still scarce; the third learns interactively, but can benefit from validation benchmarks. The development of such intelligent routing

1. A. Moussaoui is with Department of Computer Science, Intelligent Systems and Cognitive Computing (ISCC) Laboratory, University Mohamed El Bachir El Ibrahimi, Bordj Bou Arréridj, Algeria. Email: ali.moussaoui@univbba.dz
2. H. Lakhlef is with Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France.

critically depends on realistic and comprehensive datasets [11]-[12]. However, as we detail in the following section, existing datasets either lack the multi-layer information necessary for holistic FANET analysis or rely on ground-mobility assumptions that do not translate to aerial environments.

Despite increasing research activity on UAV communications, publicly available datasets specifically dedicated to FANETs remain scarce. Existing datasets are often narrowly focused on security or intrusion detection and typically lack multi-layer information, such as routing states, mobility traces, energy evolution and QoS dynamics [13]-[14]. Other large-scale *ad hoc* or vehicular datasets rely on two-dimensional ground mobility and simplified assumptions, which limits their applicability to aerial networks [15]-[16]. Consequently, there is a clear need for a general-purpose FANET dataset that provides a holistic view of UAV communication behavior.

To address these specific gaps, the dataset introduced in this paper is designed to systematically capture the three main FANET challenges identified above, high mobility, limited energy and dynamic topology, across eight controlled scenarios. Mobility speeds range from 10 to 20 m/s, initial energy budgets from 100 to 300 J and transmission ranges from 250 to 400 m . For each scenario, the dataset provides not only QoS metrics, but also the underlying packet-level traces, routing states (OLSR) and energy evolution, enabling researchers to isolate or combine these challenging factors in reproducible experiments.

Our dataset is publicly available and can be downloaded from Zenodo using the DOI: https://doi.org/10. 5281/zenodo. 19373220 [17]. It is designed to support performance evaluation, protocol benchmarking and AI-driven routing research under varying mobility, energy and communication conditions. By making this dataset publicly available, we aim to facilitate reproducible research and provide a common reference for future studies on FANETs.

The remainder of this paper is organized as follows. Section 2 reviews related work on existing FANET and mobility datasets. Section 3 presents the motivation and scientific relevance of the dataset. Section 4 describes the dataset structure and content. Section 5 details the simulation methodology. Section 6 discusses validation and consistency checks. Section 7 outlines potential applications and Section 8 provides usage guidelines. Finally, Section 9 concludes the paper and discusses future directions.

2. RELATED WORK

In recent years, research on FANETs has received a lot of attention, especially with regard to routing optimization, performance evaluation and the application of AI for adaptive communication [11]-[12]. Numerous studies have made an effort to create or model datasets for the purpose of analyzing FANET behavior; however, the majority of these datasets are still restricted in terms of their breadth, reproducibility and variety of network scenarios.

Compared to datasets created for VANETs and ground-based *Ad Hoc* networks, publicly accessible datasets for Flying *Ad Hoc* Networks (FANETs) are still incredibly rare. Given the unique features of FANETs-high node mobility, quickly changing topology, 3D movement and aerial propagation conditions - that render current datasets inappropriate for aerial contexts, this scarcity is noteworthy. FANGHETS24 is one of the few datasets specifically created for FANETs [13]. It focuses on using early time-series classification of packet interactions to detect gray hole attacks. The dataset is useful for security research, but it can't be used for general FANET networking studies, because it only includes one type of attack and doesn't offer more comprehensive network information, like routing dynamics, link stability, MAC/PHY traces or QoS performance.

UAVIDS-2025 is another pertinent dataset that provides labeled flow records for intrusion detection in UAV swarms and covers a variety of attack types, including wormhole, flooding, Sybil and black hole [14]. The dataset, which was created using NS-3 and realistic mobility, is helpful for assessing IDS models, but is still solely focused on security. Energy consumption, routing evolution, multi-protocol comparisons and low-layer communication logs are not included. Similar to FAN-GHETS24, it concentrates on a particular issue (intrusion detection) rather than offering a flexible dataset for QoS prediction, AI-based routing or FANET performance analysis.

A number of datasets on generic *Ad Hoc* networks are available outside of FANETs, such as Packet Time Delivery on *Ad Hoc* Networks by Rocha and Gradvohl [15]. Ninety thousand simulations covering packet delivery times in different node densities, regions and gateway configurations are included in

this dataset. Despite having a wide simulation coverage, it is still predicated on fixed terrestrial *Ad Hoc* assumptions and 2D mobility, lacking the mobility dynamics, aerial channel characteristics and real-time topology variations present in FANETs. Its metrics do not capture multi-layer behaviors pertinent to UAV networks and instead concentrate on delivery delays and failures.

A larger body of datasets originates from VANETs, which, although technologically distinct, illustrate how large-scale open datasets have successfully shaped research in mobile *Ad Hoc* networking. The CN+ dataset, based on real-world vehicle mobility at a signalized intersection, provides a large amount of empirical data for traffic-aware communication studies [18]. However, its constraints-ground mobility, 2D movements and traffic-light-regulated behavior-make it unsuitable for aerial networks. Similarly, the VANET Mobility Dataset integrates real highway mobility traces generated with SUMO and validated against real traffic databases (PeMS) [19]. While influential for mobility and topology studies, its channel model and movement patterns do not translate to 3D free-space UAV mobility.

Another widely used contribution is the Cologne vehicular mobility dataset, which offers 24 hours of synthetic, yet highly realistic, car mobility traces over a 400 km^2 urban area [16]. By modeling both macroscopic traffic flows and microscopic driving behavior, it has significantly improved the realism of VANET simulations. Nevertheless, despite its scale and level of detail, this dataset is limited to terrestrial mobility. It does not include communication traces, aerial dynamics, energy constraints or multi-hop routing information, which restricts its applicability to FANET-related studies. Similarly, the VeReMi dataset has become a well-established benchmark for misbehavior detection in VANETs, providing labeled benign and malicious messages in urban driving scenarios [20]. Although methodologically robust-particularly in its definition of ground truth and attack models-it remains focused on vehicle-to-vehicle safety communications and does not address aerial networking or routing-layer behavior.

Table 1 summarizes the limitations of existing datasets compared to the proposed FANET dataset.

Table 1. Comparison of existing datasets and their limitations.

Dataset	Scope & Coverage	Data Granularity	Limitations	Reproducibility & Accessibility	Key Use Cases
FANGHETS24 [13]	one scenario, grey hole attack	Packet interactions only	Security only, no QoS/routing/ energy	Available	Grey hole detection
UAVIDS2025 [14]	Multiple attacks	Flow records (IDS)	Security only, no low layer traces	NS-3 based	Intrusion detection
Packet Time Delivery [15]	90k simulations, 2D	Packet delivery times	2D, terrestrial, no aerial channel	Available	Delay analysis
Cologne [16]	24 h mobility, 2D	Mobility only	No comm./energy/ routing traces	Public	Vehicular mobility
VeReMi [20]	Urban driving	V2V messages	V2V only, no aerial multihop	Public	Misbehavior detection

Taken together, these datasets reveal two important observations. First, publicly available FANET datasets remain extremely scarce and are often narrowly tailored to specific security use cases, such as intrusion or attack detection. None of the existing datasets provides a holistic, multi-layer view of FANET behavior that jointly captures three-dimensional mobility, routing protocol states, physical-layer effects, link quality variations, QoS metrics and energy dynamics. Second, while VANET datasets clearly demonstrate the scientific value of well-designed mobility and communication traces, their underlying assumptions-ground-constrained motion, stable connectivity patterns and vehicular traffic models -are fundamentally different from those of aerial networks. As a result, they cannot be directly reused for realistic FANET modeling or AI-driven networking research.

These limitations, summarized in Table 1, highlight the need for a general-purpose, AI-ready FANET dataset that reflects the unique characteristics of aerial networks, including realistic UAV mobility, multihop routing dynamics, PHY/MAC interactions and detailed performance metrics. Such a dataset would not only support traditional networking studies, but would also enable machine learning-based applications, such as QoS prediction, routing optimization, anomaly detection, topology evolution

forecasting and autonomous swarm coordination. In contrast to existing datasets, the dataset proposed in this work is designed to address this gap by providing a comprehensive, structured and reproducible resource specifically tailored to FANET research.

3. MOTIVATION FOR DATASET CREATION AND RESEARCH SIGNIFICANCE

The increasing complexity of FANETs calls for reproducible and well-documented datasets that enable systematic evaluation of communication protocols and intelligent networking strategies. In practice, the lack of shared data sources makes it difficult to compare results across studies or to assess the robustness of proposed solutions under diverse operating conditions, leading many contributions to rely on custom simulation setups that are hard to reproduce or extend [21]-[22].

In network and artificial-intelligence research, datasets may originate from real-world measurements, such as UAV flight experiments and onboard network logs or from simulation-based environments using tools like NS-3 or OMNeT++, which allow controlled, parameterizable and repeatable experimentation. For FANETs, where real-world testing remains costly and technically challenging, simulation-based datasets represent an essential first step and can later be complemented by real-flight measurements to improve model generalization. From an AI and machine-learning perspective, the availability and quality of datasets directly affect model training, evaluation and generalization, as they provide the ground truth required for intelligent routing, QoS prediction (e.g., delay, packet-delivery ratio and energy consumption), failure detection and the design of mobility-aware and energy-efficient communication strategies. Without shared datasets, researchers are forced to repeatedly recreate similar experimental environments, resulting in poor reproducibility and inconsistent benchmarking, which ultimately hinder progress in AIdriven FANET research. In this context, datasets play a central role in the research cycle of intelligent FANET systems, serving as the foundation for data analysis, model development, performance evaluation and the continuous improvement of routing and communication protocols.

In summary, artificial-intelligence methods can be broadly divided into two categories: data-driven approaches, such as supervised and unsupervised learning and interaction-based approaches, such as reinforcement learning [24]. Data-driven approaches are the most widely used due to their strong performance; however, their effectiveness depends on the availability of high-quality datasets. This is precisely the gap that the present work addresses by providing a realistic, reproducible FANET communication dataset.

4. THE PROPOSED FANET DATASET

After discussing the general motivation and scientific significance of datasets in AI and networking research, this section focuses specifically on the proposed dataset entitled "FANET Dataset: UAV Communication Scenarios in NS-3.40". This dataset has been designed to provide a reproducible, structured and extensible resource for analyzing UAV communication behavior, energy evolution and QoS under diverse operating conditions. The following sub-sections describe in detail its structure, content and organization. The dataset is publicly available and can be accessed and downloaded from Zenodo using the DOI: https://doi.org/10.5281/zenodo.19373220.

4.1 FANET Network Architecture

In all scenarios of the proposed dataset, the network follows a unified architecture composed of a fixed base station and multiple UAV nodes operating within a bounded three-dimensional area (in our case, $1000\ \text{m} \times 1000\ \text{m} \times 200\ \text{m}$). Thus, when a scenario specifies a total node count, for example 11 nodes, this corresponds to 10 UAVs plus one base station. Figure 1 provides an illustrative overview of the architecture used throughout the dataset generation. The communication model supports three interaction types, each reflecting a distinct operational paradigm in FANET environments:

UAV-to-UAV (U2U): Communication occurs exclusively between UAVs. The base station plays a negligible role in this mode. This configuration is relevant in applications where the base station is distant or inaccessible, such as military operations or remote-area missions.

UAV-to-Base-Station (U2B): UAVs communicate directly with the base station whenever possible. However, UAVs located too far from the base station may rely on multi-hop relaying through other

UAVs to reach it. This mode represents centralized monitoring, control or data-collection scenarios.

Mixed Communication (Hybrid Mode): UAVs communicate both among themselves and with the base station. This hybrid mode corresponds to collaborative missions where UAVs coordinate locally while also transmitting situational data to a central controller.

This architectural framework ensures that the dataset captures realistic communication patterns encountered in modern FANET deployments, supporting diverse routing behaviors, mobility constraints and QoS conditions.

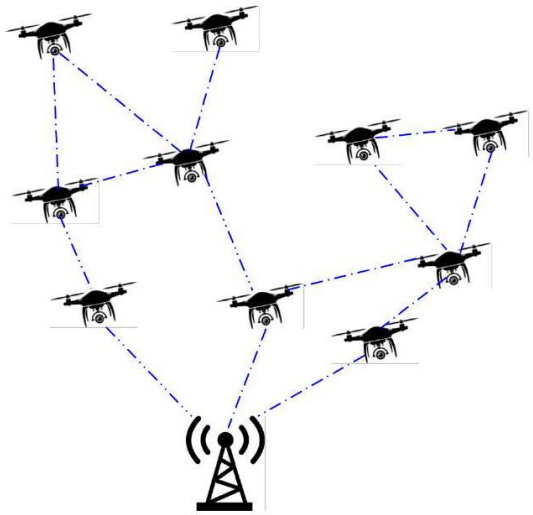


Figure 1. FANET network architecture (Example of 11 nodes, one base station and 10 drones).

4.2 Overview of Dataset Design

The FANET Dataset: UAV Communication Scenarios in NS-3.40 has been designed to provide a comprehensive, reproducible and structured source of data describing the behavior of UAV nodes operating in FANETs. The dataset captures mobility dynamics, energy consumption, routing activity and QoS metrics in various 3D network configurations. It serves as a benchmark for researchers working on routing optimization, QoS-aware communication and AI-based approaches for FANETs.

The dataset includes eight distinct simulation scenarios, each representing a specific combination of UAV density, speed, communication type and energy level. By systematically varying these parameters, the dataset provides a diverse and balanced representation of network behaviors under different environmental and operational conditions.

4.3 Scenario Structure (S1-S8)

The eight scenarios (S1 to S8) cover a range of configurations that vary in node density, mobility speed, initial energy, transmission range, traffic type and communication architecture. Table 2 provides the parameter ranges used across scenarios, while Table 3 summarizes the specific configuration of each scenario. The scenarios span from relatively stable conditions (S1: low density, low speed, high range) to more challenging ones (S6, S7, S8: high density, high speed, low range), allowing researchers to study network behavior across different regimes.

Table 2. Parameter levels for the FANET dataset scenarios.

Parameter	Level	Description / Interval
Node Density	Low Medium High	11 nodes (1 base station + 10 UAVs) 31 nodes (1 base station +30 UAVs) 51 nodes (1 base station + 50 UAVs)
Speed	Low; Medium; High	10m/s; 15m/s; 20m/s
Initial Energy	Low; Medium; High	100-150 J; 150-200 J; 200-300 J
Transmission Range	Low; Medium; High	250 m; 350 m; 400 m
Traffic Type	-	CBR / Video

Table 3. Scenario configuration summary (S1-S8).

Sc.	Architecture	Density	Speed	Energy	Range	Traffic
S1	UAV ↔UAV	Low	Low	Low	High	CBR
S2	Mixed (UAV ↔ UAV + UAV → BS)	Low	Medium	Medium	High	Video
S3	UAV → BS	Medium	Medium	Medium	Medium	CBR
S4	Mixed (UAV ↔UAV + UAV →BS)	Medium	Low	Low	Low	CBR
S5	UAV ↔UAV	Medium	High	High	Medium	Video
S6	UAV ↔UAV	High	Low	Medium	Low	CBR
S7	Mixed (UAV ↔ UAV + UAV → BS)	High	Medium	High	Low	Video
S8	UAV → BS	High	High	Low	Low	CBR

4.4 Data Organization and File Descriptions

The FANET dataset is organized in a hierarchical structure to facilitate navigation, analysis and reproducibility. Each scenario (S1-S8) is stored in a separate directory containing all CSV (Comma-separated values) files documenting the simulation results. The root directory includes the main documentation and metadata files. This organization ensures that each scenario is self-contained and can be independently analyzed, reproduced or extended.

Table 4 summarizes the main dataset files and their contents.

Note on packet counters:

- packet_trace.csv: all packets (DATA + OLSR); multi-hop retransmissions appear as separate reception events.
- olsr_links.csv: only OLSR HELLO messages (broadcasts).
- network_qos_metrics.csv and node_qos_metrics.csv:
  - sent_pkts, recv_pkts, throughput: all packets.
  - goodput, avg_delay_ms, jitter_ms: only DATA packets at final destination.
  - PDR, ETX: only DATA packets.
  - "Sent" counter: original transmission only (intermediate hops not counted).
- node_state.csv, olsr_node_state.csv: not affected by packet counters.

In addition to the CSV files described above, the Zenodo repository contains the NS-3 simulation source files that were modified or created for this work, specifically FANET_Dataset.cc, olsr-routing-protocol.cc, olsr-routing-protocol.h, ipv4-13-protocol.cc and ipv4-13-protocol.h. These files, along with the standard NS-3.40 source code, allow researchers to fully reproduce the simulations. Python postprocessing scripts used to generate the final CSV outputs are also included. A README file provides instructions for compiling and running the simulations.

Table 4. Dataset files and descriptions.

File name	Description
packet_trace.csv	Contains all transmission and reception events for each packet, including timestamps, source and destination node IDs, RSSI, SNR and delay.
network_qos_metrics.csv	Aggregated QoS metrics at the network level, including throughput, jitter, delay, packet delivery ratio, loss rate and ETX.
node_qos_metrics.csv	Node-level QoS indicators showing per-node performance and communication efficiency.
node__state.csv	Describes each UAV's energy status, position, velocity and motion parameters during the simulation.
olsr_links.csv	Details of link distances and symmetry between neighboring UAVs.
olsr_node_state.csv	Routing information related to OLSR, including neighbor sets, MPR configurations and control message counts.
simulation_scenario.csv	Metadata describing the simulation configuration, such as density level, speed, energy, transmission range, traffic type and communication architecture.

4.5 Metrics and Parameters Captured

The dataset records multiple key metrics and parameters for each UAV and the network as a whole.

- Mobility parameters: 3D positions (X, Y, Z), velocities, yaw/pitch angles.
- Energy metrics: Initial energy, remaining energy, energy consumption over time.
- QoS metrics: Throughput, delay, jitter, packet delivery ratio, loss rate, ETX, goodput.
- Routing parameters: OLSR neighbor sets, MPR sets, hello/TC message counts, link symmetry.
- Traffic characteristics: Type (CBR/video), sent/received packets and bytes per node.

The following definitions clarify how the QoS metrics are computed.

Metric definitions:

- PDR (Packet Delivery Ratio) = DestinationRecvDataPkts / sent_data_pkts (DATA packets received at final destination / DATA packets sent by the sources)
- ETX (Expected Transmission Count) = sent_data_pkts / DestinationRecvDataPkts
- Throughput (bps) = $(\text{recv\_bytes\_all} \times 8) /$ duration (all received packets, including OLSR)
- Goodput (bps) = $(\text{DestinationRecvDataBytes} \times 8) /$ duration (only DATA bytes received at final destination)
- Jitter (ms) = standard deviation of Link_Delay_ms for DATA packets
- Loss Rate = 1 - PDR

4.6 Dataset Statistics

We ran eight simulation scenarios, each lasting 200 seconds. Table 5 gives a quick overview of what each scenario generated in terms of packet traffic and file sizes. For each scenario, we included the total number of packets sent and received, the resulting packet delivery ratio (PDR), how many active flows were present and how many rows you'll find in the main CSV files (packet traces, node states and OLSR links).

Table 5. Dataset summary statistics per scenario.

Scenario	Total Packets Sent	Total Packets Received	PDR (%)	Number of Active Flows	Packet Trace Rows	Node State Rows	OLSR Links Rows
S1	52478	42529	81.0	5	128266	3723	12339
S2	56546	32061	56.7	6	130628	4389	13875
S3	65233	24700	37.9	10	252455	12344	85961
S4	71559	35422	49.5	9	213739	10251	40042
S5	52038	28657	55.1	10	230119	12369	80120
S6	56258	17428	31.0	8	300184	19838	102302
S7	49867	19275	38.7	10	265616	20349	106344
S8	53849	24491	45.5	7	230545	14659	72593

Looking at the whole dataset, we end up with roughly 2.57 million rows spread across all CSV files. That breaks down to about 1.75 million packet-related events, 0.10 million entries recording node states and 0.51 million rows of OLSR link information. The total size of the dataset is around 250 MB.

4.7 Example of Recorded Variables

To illustrate the types of data captured in the FANET Dataset, we present excerpts from two representative files: network_qos_metrics.csv and packet_trace.csv. Table 6 shows aggregated QoS metrics over consecutive time windows, such as packet delivery ratio (PDR), expected transmission count (ETX) and average link delay. Table 7 gives a closer look at individual packet events, timestamps, source and destination nodes and signal-to-noise ratio (SNR), revealing the fine-grained behavior of UAV communications.

Table 6. Excerpt from network_qos_metrics.csv showing network-level QoS metrics over time windows.

Window Start (s)	Window End (s)	Sent (DATA)	Recv (DATA)	Avg Link Delay (ms)	ETX	PDR
51.0	54.0	1438	795	0.922	1.809	0.553
54.0	57.0	1335	892	1.019	1.497	0.668
57.0	60.0	1234	655	1.350	1.884	0.531
60.0	63.0	1420	1225	1.290	1.159	0.863
63.0	66.0	1260	1113	1.370	1.132	0.883

Table 7. Excerpt from packet_trace.csv showing per-packet transmission and reception events.

TxTime (s)	PacketUid	NodeIdTx	NodeIdDst	RxTime (s)	NodeIdRx	SNR (dB)
50.1216	8890	7	Broadcast	50.1219	9	17.70
50.1216	8890	7	Broadcast	50.1219	0	16.43
50.1511	8896	3	0	50.1517	0	13.68
50.1532	8896	0	8	50.1538	8	15.88
50.1500	8898	3	Broadcast	50.1502	1	28.60

5. Methodology and Simulation Environment

5.1 NS-3.40 Setup and Simulation Configuration

All simulation scenarios were implemented using the NS-3.40 network simulator. The simulation environment represents a three-dimensional space of $1000 \times 1000 \times 200$ meters, in which UAVs follow the Random Waypoint mobility model. Each scenario runs for 200 seconds, during which all events-such as packet transmissions, receptions and node states-are recorded in CSV files to ensure reproducibility.

Table 8 summarizes the fixed simulation parameters and protocol settings used across all scenarios.

Table 8. Fixed simulation parameters and protocol settings.

Category	Parameter	Value
Wireless / PHY	Wi-Fi standard	802.11b
	Carrier frequency	2.4 GHz (default for 802.11b)
	Channel width	22 MHz (default for 802.11b)
	Propagation model	FriisPropagationLossModel
	Fading model	None
	Transmission power	Varies per scenario (12.6-16.7 dBm)
	Receiver sensitivity	-90 dBm (default)
	MAC retry limit	7 (default)
	Queue size (per interface)	100 packets (default)
Traffic	CBR traffic	Packet size $= 512$ bytes, CBR rate $= 2$Mbps
	Video traffic	VBR ($2 - 6$Mbps), packet size 1024 bytes, exponential on/off
Mobility	Mobility model	3D Random Waypoint
	Pause time	10 s
	Waypoint generation	Uniform random within simulation area ($0 - 1000$ m in X, $0 - 1000$ m in Y, $0 - 200$ m in Z)
	Altitude range	$0 - 200$ m
	Speed min/max	As per scenario ($5 - 20$ m/s)
Simulation	Simulation duration	200 s
	Random seeds	Default ns-3 seeds (not explicitly set)
	Number of repetitions	1 per scenario (no repeated runs)

5.2 Routing Protocol

We chose OLSR as the routing protocol across all scenarios. Its proactive nature means it keeps routing tables updated continuously rather than waiting for a route request-something we found useful in highly dynamic environments, like FANETs. The standard OLSR implementation in ns- 3 was used without modifying its internal logic. We did, however, adjust two timing parameters to better suit UAV mobility: the HELLO interval was lowered from 2s to 0.5 s and the TC interval from 5s to 1.5s. This allows neighbor detection and topology updates to happen more frequently, which is helpful when nodes move quickly in 3D space.

It's worth noting that the goal here isn't to evaluate OLSR itself or claim it's the best choice for FANETs. Instead, we use it as a stable baseline to collect rich data-mobility traces, link changes, neighbor relationships and routing events, like HELLO messages, MPR selections and TC updates-all saved in CSV format. Our hope is that the community can use this dataset to experiment with their own routing protocols, whether traditional, AI-based or QoS-aware, using realistic aerial mobility and communication traces.

5.3 Mobility and Energy Models

We modeled UAV mobility in three dimensions using the random waypoint model, with minimum and maximum speed variations (10, 15 and $20\ \mathrm{m/s}$) according to scenario settings. Each UAV node includes an energy source (BasicEnergySource) and a radio energy model (WifiRadioEnergyModel) to simulate energy consumption for transmissions and receptions. Node positions, velocities orientations (yaw, pitch) and remaining energy are logged at each time step.

5.4 Parameter Space

The dataset explores a multi-dimensional parameter space to capture diverse network behaviors:

- **Node density:** Low (11 total nodes: 1 base station +10 UAVs), Medium (31 total nodes: 1 base station +30 UAVs), High (51 total nodes: 1 base station +50 UAVs)
- **Speed:** 10, 15, $20\ \mathrm{m/s}$
- **Transmission range:** $250\ \mathrm{m}, 350\ \mathrm{m}, 400\ \mathrm{m}$
- **Traffic type:** CBR and video streams
- **Communication types:** UAV ↔UAV, UAV → BS or Mixed

These parameters are systematically varied across the eight scenarios to ensure a comprehensive dataset for AI-based routing studies, QoS evaluation and energy-aware network analysis.

5.5 Experimental Reproducibility

To ensure reproducibility, all scenarios use explicitly defined random seeds for mobility and traffic generation. Simulation logs record every packet event, node state and QoS metric with time stamps. Scenariospecific configuration files document the parameter values, allowing any researcher to exactly reproduce a given simulation or extend it with additional UAVs or traffic conditions.

5.6 Dataset Construction Workflow

The proposed FANET dataset is generated through a multi-stage pipeline that combines physical-layer traces, routing-layer logs, node mobility and energy states and Python-based post-processing for QoS metric extraction. Each final CSV file included in the dataset corresponds to a specific layer or functional aspect of the FANET communication process.

Physical Layer Packet Traces: At the PHY layer, NS-3.40 is instrumented to record all transmission and reception events. Two raw data streams are produced:

- tx.csv: captures every transmission event (Phy/TxBegin), including transmission time, packet UID, transmitting node ID, power level, packet size and data rate.
- rx.csv: records each reception event (Phy/RxEnd), including reception time, receiving node ID, RSSI, noise, SNR and link delay.

These two sources are combined to create the final file:

- packet_trace.csv: a unified table linking each transmitted packet to its corresponding receptions. It includes PHY-layer metrics, such as RSSI, SNR, noise power, transmission power, link delay, packet type classification and sender/receiver roles.

This file represents the complete physical communication behavior of the FANET.

Routing Layer Traces (OLSR): The OLSR routing protocol is instrumented to extract protocol-level state evolution:

- olsr_links.csv: lists, for each node and each time step, its neighbor nodes, link symmetry status and inter-node distance.
- olsr_node_state.csv: records OLSR internal dynamics including MPR selection, one-hop and two-hop neighbor counts, HELLO and TC activity and MPR set evolution.

These files provide a protocol-centric perspective on network topology, stability and routing behavior.

Node Mobility, Energy and Traffic State: A separate trace provides detailed node-level evolution over time:

- node_state.csv: includes mobility information (3D position, velocity, speed, yaw, pitch), energy status (initial and remaining energy) and traffic parameters (data mode and data rate).

This file is essential for correlating communication performance with mobility patterns and energy constraints of UAVs.

QoS Metrics Generation via Python Post-processing: Python scripts process the raw physical and routing logs to compute network-wide and per-node performance metrics:

- network_qos_metrics.csv: aggregates global metrics over time windows, including sent and received packets/bytes, average delay, jitter, throughput, goodput, ETX, Packet Delivery Ratio (PDR) and packet loss rate.
- node_qos_metrics.csv: provides per-UAV QoS indicators, such as per-node throughput, delay, jitter, sent/received data, delivered data packets and goodput.

These tables enable fine-grained evaluation of performance at both local (node-level) and global (networklevel) scales.

Scenario Description: Finally, a scenario descriptor summarizes all configuration parameters:

- simulation_scenario.csv: defines density levels, mobility categories, energy configurations, communication range levels, traffic type, communication patterns (UAV-to-UAV, UAV-to-BS or mixed), number of source/destination UAVs, protocol (OLSR), total duration and simulation zone size.

Final Dataset Organization: All tables are generated with consistent timestamps, node identifiers and unified variable naming. Together, the dataset provides:

- physical-layer communication traces,
- routing-layer protocol dynamics,
- UAV mobility and energy evolution,
- QoS metrics at network and node levels,
- full scenario descriptors for reproducibility.

This multi-layer architecture offers a comprehensive view of FANET behavior across diverse configurations and enables cross-layer analysis for UAV communication research.

6. Dataset Validation and Consistency Checks

6.1 Data Verification

To ensure the reliability and scientific integrity of the FANET dataset, multiple validation and consistency checks were performed. Each CSV file was examined for:

- Missing values: All fields are verified to contain valid entries. For instance, packets that were not received are explicitly marked with NaN values in RxTime_s, NodeIdRx or RSSI.

- Consistency across runs: Metrics, such as total sent and received packets, node energy evolution and connectivity patterns, are compared across multiple simulation repetitions using different random seeds.
- Range validity: All parameters are checked against expected ranges (e.g., UAV speeds between 10-20 m/s, energy levels within initial configuration bounds, transmission power within radio model specifications).
- Protocol integrity: OLSR routing variables, such as neighbor sets, MPR selections and TC messages, are verified to align with network topology changes over time.

In addition to these internal checks, we examined how key QoS metrics evolve across the eight scenarios to verify that the dataset behaves consistently with expected network dynamics. Figure 3 shows the average Packet Delivery Ratio (PDR) and Figure 2 shows the average Expected Transmission Count (ETX) for each scenario, with error bars indicating the standard deviation over time. For these averages, windows without any DATA traffic were excluded to avoid biasing the metrics.

As expected, Scenario 1 (low mobility, high transmission range) achieves the highest PDR (0.848) and the lowest ETX (1.277), indicating very reliable communication. In contrast, Scenarios 6 and 7 (high node density, low transmission range) exhibit the lowest PDR (0.375 and 0.305) and the highest ETX (13.06 and 24.30), reflecting the increased link instability and congestion under challenging network conditions. Video traffic (Scenarios 2, 5 and 7) results in higher throughput, but lower goodput, due to the increased load, which is consistent with the behavior of real FANET deployments. These trends confirm that the dataset captures realistic and coherent performance variations across different operational conditions.

Together, these checks confirm that the dataset reflects physically plausible UAV behavior and network dynamics, making it well-suited for research applications, such as AI-based routing, QoS evaluation and energy-aware FANET protocol design.

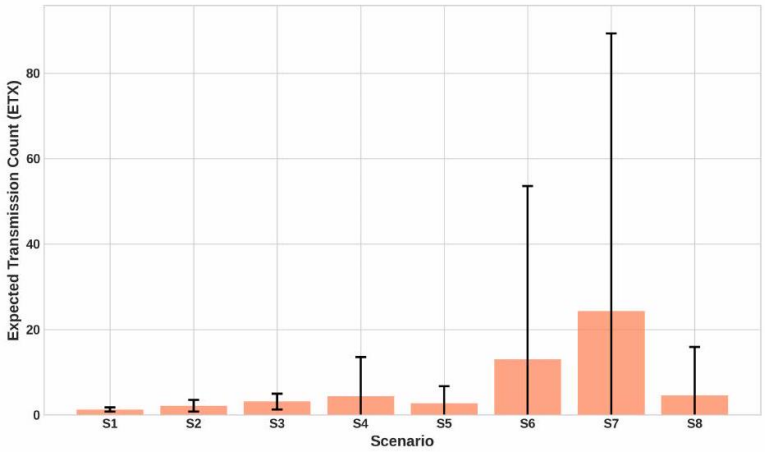


Figure 2. Average expected transmission count per scenario.

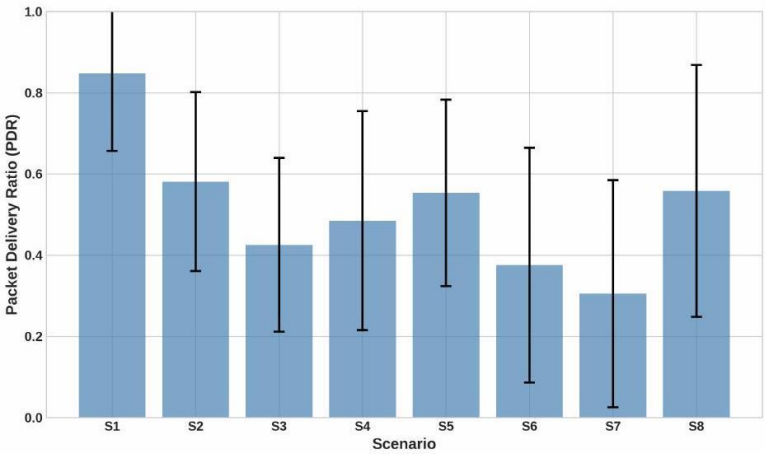


Figure 3. Average packet delivery ratio per scenario

6.2 Scientific Integrity

By systematically performing these checks, the dataset ensures:

- No internal contradictions between node-level and network-level metrics.
- Reproducibility across simulation runs.
- Accurate representation of UAV mobility, energy consumption and QoS performance.

These measures provide confidence that the dataset is suitable for research applications, including AIbased routing, QoS evaluation and energy-aware FANET protocol design.

7. POTENTIAL APPLICATIONS AND RESEARCH IMPACT

The FANET dataset offers a rich source of information that can support multiple research directions in UAV communication and networking. Its detailed recordings of node mobility, energy consumption, QoS metrics and protocol behavior enable the following applications:

7.1 AI-based Routing Model Training

The dataset provides labeled, time-stamped data suitable for training machine-learning and artificial-intelligence models aimed at routing optimization. Researchers can use node-level and network-level metrics to develop predictive or adaptive routing strategies that account for UAV mobility, link quality and energy constraints.

7.2 Quality of Service Optimization

Aggregated network QoS metrics, such as throughput, goodput, delay, jitter, PDR and ETX, allow for the evaluation and tuning of communication protocols. This facilitates the design of QoS-aware routing and scheduling strategies that ensure reliable UAV network performance under diverse operational conditions.

7.3 Energy-aware Communication Strategies

By capturing detailed energy consumption patterns of UAV nodes over time, the dataset enables the development of energy-efficient routing and transmission schemes. This is critical for extending UAV operational time and maintaining network connectivity in multi-UAV scenarios.

7.4 Topology Prediction and Link-stability Analysis

Recorded mobility and link-state information supports the study of dynamic-topology evolution. Researchers can analyze neighbor sets, MPR selection and link symmetry to predict network connectivity and link stability, which are essential for both protocol design and real-time network management.

7.5 Relevance across UAV, FANET, MANET and IoT Research

While focused on FANETs, the dataset provides insights applicable to broader networking contexts, including MANETs, VANETs and IoT-based UAV systems. Comparative studies across these domains are facilitated by the structured and reproducible nature of the dataset.

7.6 Benchmarking and Reproducible Research

The eight scenario configurations, with diverse densities, speeds, traffic types and energy levels, enable reproducible experiments and benchmarking of new algorithms. We didn't try to cover every possible combination; instead, we picked operating points that reflect realistic trade-offs across the main factors that affect network performance: density, mobility, transmission range, energy, traffic and architecture. The scenarios go from relatively easy conditions (S1: low density, low speed, high range) to much harder ones (S6, S7, S8: high density, high speed, low range). This lets researchers see how performance changes across different regimes without drowning in too many scenarios. Having multiple parameters vary at once is intentional, it pushes machine-learning models to learn how factors interact rather than just memorize isolated cases. If someone needs a different setup or wants to isolate a specific parameter, he/she can easily generate new scenarios using the simulation scripts we provide.

Researchers can replicate or extend scenarios to evaluate novel-routing, energy-management or QoS-optimization techniques in a controlled, realistic setting.

8. USAGE NOTES

The FANET dataset is designed to be easily accessible and reusable by the research community. The following instructions and recommendations support effective utilization:

8.1 Accessing the Dataset

The dataset is organized into eight scenario directories, each containing CSV files that describe node behavior, network QoS and simulation metadata. Users can download the dataset from the provided repository or Zenodo DOI, https://doi.org/10.5281/zenodo.19373220. The hierarchical structure ensures that each scenario can be independently analyzed or reproduced.

8.2 Recommended Pre-processing

Before using the dataset for analysis or modeling, we recommend the following pre-processing steps:

- Handle any missing or NaN values, especially in packet traces, using interpolation or filtering if required.
- Normalize or scale numeric variables, such as energy levels, throughput and delays, for machine-learning applications.
- Aggregate per-node or per-window metrics for comparative analysis across scenarios.
- Convert time units or synchronize timestamps if combining multiple scenario files for simulation-wide studies.
- For machine-learning experiments, avoid data leakage by splitting data appropriately. We recommend either using time-based splits (e.g., first 80% of timesteps for training, last 20% for testing) or splitting by simulation runs (e.g., train on a sub-set of scenarios, test on unseen ones).

8.3 Integration with Analysis Tools

The dataset is designed to be easily integrated into commonly used data analysis and simulation environments. It is fully compatible with standard scientific workflows and can be processed using widely adopted tools. In particular, Python-based environments allow efficient parsing, processing and visualization of the CSV files through libraries, such as pandas, NumPy and Matplotlib.

The dataset can also be imported into MATLAB using functions, such as readtable or csvread, enabling further modeling, statistical analysis and detailed evaluation of QoS metrics. In addition, the CSV output files can be reused within the NS-3 framework to validate custom simulation runs, reproduce the original scenarios or serve as input data for training and evaluating AI-based routing and optimization models.

8.4 Example Machine-learning Task

To illustrate how the dataset can be used for machine learning, consider predicting link delay (Link_Delay_ms) from physical-layer features. The packet_trace.csv file provides RSSI, noise, SNR and delay for each packet reception. A researcher could use these features to train a model that estimates delay from signal quality, useful for applications sensitive to latency.

For more advanced tasks, the dataset allows combining information from multiple files. For example, by joining packet_trace.csv (RSSI, noise, SNR, delay) with node_state.csv (remaining energy, velocity components) on node ID and timestamp, one can build a richer feature set that includes both link quality and node state. This enables predicting delay under varying mobility and energy conditions.

To further facilitate machine-learning experiments, the Zenodo repository includes an example_of_ml_ready folder containing pre-processed files for a sample task: link-stability prediction. Each scenario (S1 to S8) has a corresponding Link_Stability_For_ML_ScenarioX.csv file, built by joining olsr_links.csv, node_state.csv and packet_trace.csv. These files contain features, such as distance, signal quality (RSSI, SNR), node mobility (speed, velocity components), remaining energy and the label link_lifetime_s, which represents the estimated link durability. Researchers may use this

pre-processed dataset directly to train models that predict link stability or they may define their own labels and feature sets from the raw CSV files.

These examples are just meant to show the dataset's flexibility, users can define their own prediction, classification or clustering tasks depending on their research goals.

8.5 Citation and License

Users are reminded to appropriately cite the dataset in any derivative work:

Ali, MOUSSAOUI; Hicham, Lakhlef (2026). FANET Dataset: UAV Communication Scenarios in NS-3.40. Zenodo. DOI: https://doi.org/10.5281/zenodo.19373220.

When the dataset or any derived version is used in a scientific work, this article should also be cited in addition to the dataset itself.

9. Conclusion and Future Perspectives

In this paper, we introduced a comprehensive dataset of UAV communication scenarios for FANETs, generated using the NS-3.40 simulator. Our dataset captures key aspects of FANET operation, including UAV mobility dynamics, energy consumption, QoS metrics and OLSR routing behavior across eight diverse simulation scenarios. By offering well-structured and fully reproducible data, the dataset provides a practical reference for researchers studying FANET performance, exploring routing optimization and developing AI-driven networking solutions. It also offers a solid data foundation for training and validating AI-based routing algorithms in three-dimensional UAV networks, while supporting reproducible experimentation and fair comparison across different FANET configurations.

A few clarifications are worth mentioning. The dataset relies on the 3D Random Waypoint mobility model and uses OLSR as the routing protocol. Both are standard choices that provide a reproducible baseline. The eight scenarios combine several parameters at once (density, speed, range, energy, traffic), which reflects the kind of trade-offs you would face in real deployments. We see these not as limitations, but as defining the scope of the current release. Since all simulation scripts are publicly available, other researchers can easily adapt the framework to include other mobility models, routing protocols or parameter setups if needed.

Looking ahead, we plan to extend the dataset in several directions. First, we will add more routing protocols, including AI-assisted approaches, to support broader benchmarking. Second, we aim to introduce more varied mobility patterns and eventually include data from real UAV flights. Third, we want to explore environmental effects and larger network scales to bring the simulations closer to real-world conditions.

Our goal is to keep enriching this resource so it can better support research on AI-driven, energy-aware communication strategies and help bridge the gap between simulation and real UAV deployments.

References

[1] I. Bekmezci, O. K. Sahingoz and Ş. Temel, "Flying Ad-hoc Networks (FANETs): A Survey," Ad Hoc Networks, vol. 11, no. 3, pp. 1254-1270, 2013.

[2] T. K. Bhatia et al., "Flying Ad-Hoc Networks (FANETs): A Review," EAI Endorsed Transactions on Energy Web, vol. 11, no. 10, 2024.

[3] A. Chriki et al., "FANET: Communication, Mobility Models and Security Issues," Computer Networks, vol. 163, p. 106877, 2019.

[4] F. Pasandideh et al., "A Review of Flying Ad Hoc Networks: Key Characteristics, Applications and Wireless Technologies," Remote Sensing, vol. 14, no. 18, p. 4459, 2022.

[5] O. S. Oubbati et al., "A Survey on Position-based Routing Protocols for Flying Ad Hoc Networks (FANETs)," Vehicular Communications, vol. 10, pp. 29-56, 2017.

[6] M. J. Almansor et al., "Routing Protocols Strategies for Flying Ad-Hoc Network (FANET): Review, Taxonomy and Open Research Issues," Alexandria Engineering Journal, vol. 109, pp. 553-577, 2024.

[7] C. A. T. Romero et al., "FANET and MANET, a Support and Composition Relationship," Computers, Materials Continua, vol. 82, no. 2, 2025.

[8] G. Amponis et al., "A Survey on FANET Routing from a Cross-layer Design Perspective," Journal of Systems Architecture, vol. 120, p. 102281, 2021.
[9] T. Clausen and P. Jacquet (eds.), "RFC 3626: Optimized Link State Routing Protocol (OLSR)," IETF, RFC Editor, United States, DOI: https://doi.org/10.17487/RFC3626, 2003.
[10] C. Perkins, E. Belding-Royer and S. Das, "RFC 3561: Ad Hoc On-demand Distance Vector (AODV) Routing," IETF, RFC Editor, United States, DOI: https://doi.org/10.17487/RFC3561, 2003.
[11] P. Khoshvaght et al., "Computational Intelligence-based Routing Schemes in Flying Ad-hoc Networks (FANETs): A Review," Vehicular Communications, vol. 53, p. 100913, 2025.
[12] M. Kaur et al., "Machine Learning-based Routing Protocol in Flying Ad Hoc Networks: A Review," Computers, Materials Continua, vol. 82, no. 2, 2025.
[13] C. Hutchins et al., "A Flying Ad-hoc Network Dataset for Early Time Series Classification of Grey Hole Attacks," Scientific Data, vol. 12, no. 1, p. 1431, 2025.
[14] Q. Zeng, A. Bashir and F. Nait-Abdesselam, "UAVIDS-2025: A Benchmark Dataset for Intrusion Detection in UAV Networks Using Machine Learning Techniques," Proc. of the 2025 IEEE Conf. on Comm. and Network Secur. (CNS), DOI: 10.1109/CNS66487.2025.11194990, Avignon, France, 2025.
[15] R. Rocha and A. Gradvohl, "Packet Time Delivery on Ad Hoc Network," [Online], Available: https://zenodo.org/record/817024, Version v1, 2012.
[16] S. Uppoor et al., "Generation and Analysis of a Large-scale Urban Vehicular Mobility Dataset," IEEE Transactions on Mobile Computing, vol. 13, no. 5, pp. 1061-1075, 2013.
[17] A. Mousaoui and H. Lakhlef, "FANET Dataset: UAV Communication Scenarios in NS-3.40," [Online], Available: https://doi.org/10.5281/zenodo.19373220, Version 2.0, 2026.
[18] T. Karunathilake, M. Zongo, D. Amarawardana and A. Förster, "CN+: Vehicular Dataset at Traffic Light Regulated Intersection in Bremen, Germany," Scientific Data, vol. 11, no. 1, p. 665, 2024.
[19] N. Akhtar, S. C. Ergen and O. Ozkasap, "Vehicle Mobility and Communication Channel Models for Realistic and Efficient Highway VANET Simulation," IEEE Transactions on Vehicular Technology, vol. 64, no. 1, pp. 248-262, 2014.
[20] R. W. Van Der Heijden, T. Lukaseder and F. Kargl, "VeReMi: A Dataset for Comparable Evaluation of Misbehavior Detection in VANETs," Proc. of the Int. Conf. on Security and Privacy in Communication Systems (SecureComm 2018), pp. 318-337, Cham: Springer International Publishing, 2018.
[21] A. Rovira-Sugranes et al., "A Review of AI-enabled Routing Protocols for UAV Networks: Trends, Challenges and Future Outlook," Ad Hoc Networks, vol. 130, p. 102790, 2022.
[22] S. Thirumuruganathan et al., "Data Curation with Deep Learning," Proc. of the 23rd Int. Conf. on Extending Database Technology (EDBT), pp. 277-286, DOI: 10.5441/002/edbt.2020.25, 2020.
[23] V. Singh et al., "A-Geo: Adaptive Geographic Routing for Consumer FANETs in Next-generation Communication," IEEE Transactions on Consumer Electronics, vol. 71, no. 4, pp. 11034-11043, 2025.
[24] K. Lv et al., "Large Language Model-empowered Energy-efficient Multi-UAV-Assisted MEC Heterogeneous Networks," IEEE Transactions on Cognitive Communications and Networking, vol. 12, pp. 5281-5294, 2025.

ملخص البحث:

تمكّن شبكات الطّائرات بدون طيار من التّواصل بين الطّائرات بدون طيار في بيئاتٍ شديدة الديناميكية تفتقر إلى البنية التّحتية. ومع ذلك، فإنّ الحركة العالية، ومحدودية الطّاقة المخزّنة، والتّغير السّريع في بنية الشّبكة تجعل من ضمان موثوقية الاتّصال وجودة الخدمة تحدّياً كبيراً.

تقدم هذه الورقة مجموعة بياناتٍ متاحة للعموم لشبكات الطّائرات بدون طيار تمّ إنشاؤها من خلال محاكاة تفصيلية باستخدام برنامج (NS-3.40)، وتتكوّن من ثمانية سيناريوهات اتّصال تتغير فيها كثافة العُقد، وسرعة الحركة، ومدى الإرسال، ومستويات الطّاقة، ونوع حركة البيانات، وبنية الاتّصال لكلّ سيناريو. وقد صُمّمت مجموعة البيانات لدعم تحليل الأداء، وقياس أداء البروتوكولات، وتطوير استراتيجيات توجيه مُراعية للطّاقة ومعتمدة على الذّكاء الاصطناعي. وهي تهدف إلى تسهيل إجراء تجارب قابلة للتّكرار، وتوفير مرجعٍ علمي عملي للبحوث المستقبلية في مجال شبكات اتّصالات الطّائرات بدون طيار.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

FOUNTAIN CODES-BASED HYBRID SATELLITE TERRESTRIAL RELAY MULTICAST SCHEMES IN CO-CHANNEL INTERFERENCE ENVIRONMENT: OUTAGE PERFORMANCE, JOINT TIME AND POWER ALLOCATIONS

Nguyen Van Toan[1], Nguyen Ngoc Lan[2], Tran Trung Duy[2*], Pham Ngoc Son[3] and Nguyen Trung Hieu[2]

(Received: 1-Feb.-2026, Revised: 2-Apr.-2026 and 29-Apr-2026, Accepted: 29-Apr.-2026)

ABSTRACT

In this paper, we study outage performance of hybrid satellite-terrestrial relay multicast schemes employing Fountain codes. In the considered schemes, a satellite attempts to transmit its data to a group of ground users with the assistance of a terrestrial relay station. In the conventional scheme (referred to as ConV), the relay station forwards each Fountain packet to the ground users using decode-and-forward (DF). In the proposed scheme (referred to as ProP), the relay station stores Fountain packets received from the satellite and replaces the satellite in transmitting new Fountain packets to the ground users once it has collected a sufficient number of Fountain packets for data recovery. We derive exact closed-form expressions of outage probability (OP) at each user and system outage probability (SOP) for the ConV and ProP schemes, considering the impact of co-channel interference. Computer simulations are realized to validate the derived formulas. Moreover, a joint time and power allocation problem is formulated and solved to optimize the SOP performance for the two considered schemes.

KEYWORDS

Hybrid satellite-terrestrial relay multicast networks, Fountain codes, Co-channel interference, Outage probability, Joint time and power allocation.

1. INTRODUCTION

This work considers Hybrid Satellite-Terrestrial Relay (HSTR) schemes, in which terrestrial relays are utilized to facilitate communication between satellites and end users on the ground [1]-[3]. By integrating these relay stations, the HSTR schemes benefit from improved signal robustness, broader service areas, and greater reliability under challenging channel conditions. Such schemes are envisioned as a key enabler for next-generation infrastructures, offering high-speed access and minimal latency to meet future connectivity requirements. In [2], high-altitude platforms (HAPs) were presented as a complementary component to satellite networks within hybrid architectures, offering improved efficiency and addressing performance gaps in satellite-driven services, such as data relaying and fleet coordination. In [3], the authors proposed adaptive transmission schemes for HSTR networks to improve spectral and power efficiency in practical applications. Reference [4] introduced unmanned aerial vehicle (UAV)-aided maritime communication networks to enhance coverage and performance, acting as a supplementary layer to marine satellites and shore-based terrestrial stations. In [5], a space-air-ground free-space optical (FSO) network incorporating a high-altitude relay was proposed to enhance the reliability of satellite communications. The works in [6]-[10] have focused on system-level optimization and design for dynamic satellite-terrestrial integrated networks, including network function placement, beam direction control, radio resource allocation, multicast transmission, and fog/edge computing architectures, with the objective of enhancing communication performance and quality of service. Recent studies, such as [11], have further explored advanced communication-computation co-design frameworks for integrated satellite and aerial networks. However, these studies mainly address system-level aspects rather than physical-layer performance analysis under practical impairments, such as co-channel interference. In [12]-[13], the authors evaluated secrecy performance of the HSTR

1. N. V. Toan is with Ho Chi Minh City University of Technology and Engineering, Ho Chi Minh City, Vietnam, and Telecommunications University, Nha Trang City, Vietnam. Email: toannv.ncs@hcmute.edu.vn
2. N. N. Lan, T. T. Duy (Corresponding Author) and N. T. Hieu are with Posts and Telecommunications Institute of Technology, Ha Noi, Vietnam. Emails: lannn@ptit.edu.vn, duytt@ptit.edu.vn and hieunt@ptit.edu.vn
3. P. N. Son is with Ho Chi Minh City Uni. of Technology and Engineering., Ho Chi Minh City, Vietnam Email: sonpndtvt@hcmute.edu.vn

schemes in the presence of eavesdroppers. In addition, Reference [12] examined a scenario involving multiple eavesdroppers, while Reference [13] proposed relay selection strategies. The authors of [14] investigated the trade-off between intercept probability (IP) at the eavesdropper and outage probability (OP) at legitimate receiver for the HSTR models. Furthermore, an effective relay selection mechanism was proposed in [14] to improve the IP/OP trade-off under the impact of co-channel interference (CCI). Researchers in [15]-[17] investigated the HSTR models operating in cognitive radio (CR) environments, where the licensed spectrum owned by primary users can be shared with secondary users, as long as operation of secondary users was not harmful to performance of primary users. Another research direction involves the use of wirelessly energy harvesting in the HSTR schemes [18]-[19]. Indeed, the wireless devices in [18]-[19] have to harvest energy from surrounding radio-frequency signals. In [20]-[21], full-duplex relay techniques have been investigated in the context of the HSTR schemes, enabling simultaneous data transmission and reception at relay nodes equipped with multiple antennas. Reference [22] introduced an HSTR scenario supported by reconfigurable intelligent surfaces (RIS). Unlike the conventional relaying approaches in which relay nodes actively process signals, RIS-based relaying models utilize intelligent reflective elements to direct incoming wireless signals toward intended destinations in an optimized manner. In [23]-[24], the authors integrated Non-Orthogonal Multiple Access (NOMA) into the HSTR systems, enabling the satellite to simultaneously transmit different data to multiple ground users. In [23], the authors proposed and derived expressions of OP for the NOMA -assisted HSTR systems. Reference [24] further considered the impact of direct communication links between the satellite and the ground NOMA users. The OP performance of the NOMA -based HSTR schemes operating in CR environments was evaluated in [25]. Reference [26] investigated both IP and OP performance for cognitive users in the NOMA -aided HSTR models. In [27], OP of multi-relay NOMA -based HSTR networks was also derived and validated. In contrast to the aforementioned studies, this paper considers the HSTR scheme that incorporates Fountain codes (FCs).

Fountain codes (FCs) [28]-[29] have demonstrated effectiveness in wireless networks, thanks to their ease of implementation and ability to adapt to changing environmental conditions. Recently, several studies [30]-[34] have reported on the HSTR models employing FCs. In [30], the OP performance of the HSTR model employing FCs under the CCI condition was evaluated. Reference [31] studied the OP/IP trade-off of the FCs -based HSTR schemes employing the artificial jamming technique to reduce quality of the eavesdropping links. In [32], both NOMA and RIS were integrated into the FC-aided HSTR systems to improve secrecy rate throughput, with the presence of multiple eavesdroppers. The authors of [33] investigated OP of joint NOMA and FC-based HSTR scenarios incorporating two groups of ground users. Published works [34] studied the OP/IP trade-off for the NOMA -aided HSTR multicast schemes using FCs and a partial terrestrial station selection algorithm.

To highlight the contribution of this work, Table 1 provides a comparative summary of representative FC-aided HSTR studies in terms of transmission scenario, interference modeling, performance metrics, and optimization capability.

Table 1. Summary and comparison of related works.

Ref.	Transmission Scenario	FC Used	CCI at Relay	CCI at Users	Performance Metrics	Time - Power Allocation
[30]	HSTR relaying (broadcast)	✓	*X*	✓	OP, SOP	*X*
[31]	STN with friendly jamming	✓	*X*	*X*	OP, IP	*X*
[32]	Multi-user HSTRN (NOMA + IRS)	✓	*X*	*X*	OP, IP	*X*
[33]	HSTR broadcast (NOMA-based)	✓	*X*	*X*	OP, SOP	*X*
[34]	HSTR multicast (NOMA + PRS)	✓	*X*	*X*	OP, IP, SOP, SIP	*X*
Our work	HSTR multicast with CCI	✓	✓	✓	OP, SOP	✓

As observed from Table 1, none of the existing studies jointly considers multicast transmission, dual CCI effects at both relay and users, and joint time-power allocation. More importantly, the interplay among these factors introduces several fundamental challenges beyond a straightforward combination of existing techniques.

Specifically, multicast transmission inherently couples the decoding performance of all users, making the system outage probability depending on the worst-channel condition, which significantly complicates the analysis compared to conventional unicast or broadcast scenarios. In addition, the presence of dual CCI at both the relay and user sides makes the end-to-end outage event jointly dependent on multiple transmission phases, thereby preventing the direct use of simplified independent-link analysis commonly adopted in prior studies. Furthermore, the joint optimization of time and power allocation under multicast and interference-limited conditions results in a highly coupled and non-convex problem, where the trade-off between reliability and interference mitigation becomes significantly more intricate. Therefore, the proposed framework should be regarded as a fundamentally new system model rather than a simple extension of existing works.

Motivated by these research gaps, this paper investigates FC-aided HSTR schemes, where a satellite communicates with a group of ground users *via* a terrestrial relay station. Unlike [31]-[34], our proposed schemes take into account the impact of CCI on the OP performance. In contrast to [30], this study considers the effect of CCI on both the relay station and all ground users. Furthermore, two FC-aided HSTR schemes are considered; namely, the conventional forwarding scheme (ConV) and the proposed packet-accumulation-based scheme (ProP), to address the identified research gaps. Compared with the ConV scheme, the proposed ProP scheme improves transmission reliability by exploiting the packet-accumulation property of FC. This design is particularly suitable for interference-limited or poor channel conditions, where packet-by-packet forwarding may become inefficient. For performance measurement, exact closed-form expressions of OP at each ground user and system outage probability (SOP) are derived for ConV and ProP. The accuracy of the analytical results is verified through computer simulations. Finally, a joint optimization problem involving time and power allocation is formulated and solved to enhance the SOP performance for the two considered schemes.

The main contributions of this paper can be summarized as follows:

- Development of an FC-aided HSTR multicast model under dual CCI affecting both relay and user nodes.
- Design of a packet-accumulation-based ProP scheme, in which the relay collects sufficient Fountain-coded packets before forwarding to improve transmission reliability.
- Derivation of exact closed-form expressions for the OP at each user and the SOP for both ConV and ProP schemes.
- Joint optimization of time and power allocation to minimize SOP using an efficient GSS-based approach.

The remainder of this article is organized as follows: Section 2 describes the system model of the proposed schemes. Section 3 derives OP and SOP of ConV and ProP. Section 4 provides Monte-Carlo simulations to validate the analytical formulae. Finally, Section 5 concludes the paper.

2. SYSTEM MODEL

Figure 1 presents the system model of the proposed FC-aided HSTRNs, with presence of K interference sources. In particular, a satellite (S) tries to send the same data w_{s} to M ground users. Let us denote the ground users as $\mathrm{U}_m(m = 1,2,\dots,M)$ and the interference sources as $\mathrm{I}_k(k = 1,2,\dots,K)$. It is assumed that there is no direct link between S and U_m, as we consider a worstcase scenario in which ground users experience blockage, severe shadowing, or significant path loss (e.g., urban canyon environments, indoor users, or obstructed areas), which are commonly encountered in practice and render the satellite-user link highly unreliable [13], [35]; therefore, the data transmission between S and U_m is assisted by a terrestrial relay station (R).

In the considered model, co-channel interference affects both decoding stages; namely the satellite-to-relay reception and the relay-to-user reception. Specifically, the relay is interfered by the links $\mathrm{I_k} \rightarrow \mathrm{R}$, while each user U_m is interfered by the links $\mathrm{I_k} \rightarrow \mathrm{U_m}$. We consider a normalized two-slot transmission framework with fixed slot duration and assume perfect synchronization between the satellite and the relay, which is consistently applied to both the ConV and ProP schemes.

Using FCs, S generates encoded packets (denoted as p_{S}) from the original data (w_{S}), and these encoded packets are then transmitted from S to U_m *via* the help of R , using the DF approach. In order to successfully recover the original data $w_{\mathrm{S}}, \mathrm{U}_m$ needs to collect at least $G_{\min}$ encoded packets p_{S}, where

$G_{\min}$ represents the minimum number of packets required for FC decoding and is typically given by $G_{\min} = P(1+\varepsilon)$. Here, P denotes the number of source packets and ε is the small FC overhead that ensures reliable recovery (generally $\in [0.02,0.1]$) [29]-[31]. In addition, the satellite is allowed to transmit at most $N_{\max}$ encoded packets due to the delay constraints of the HSTR system. Since successful decoding is impossible if fewer than $G_{\min}$ packets are transmitted, the condition $N_{\max} \geq G_{\min}$ must always hold to guarantee recoverability at R and all ground users. These definitions help clarify the FC parameter selection used in the proposed multicast schemes. All nodes $\mathrm{S}, \mathrm{R}, \mathrm{U}_m$ and I_k are assumed to be equipped with a single antenna. This simplified model facilitates tractable analysis while still capturing the essential behavior of the considered system. Extensions to multiantenna scenarios will be investigated in future work.

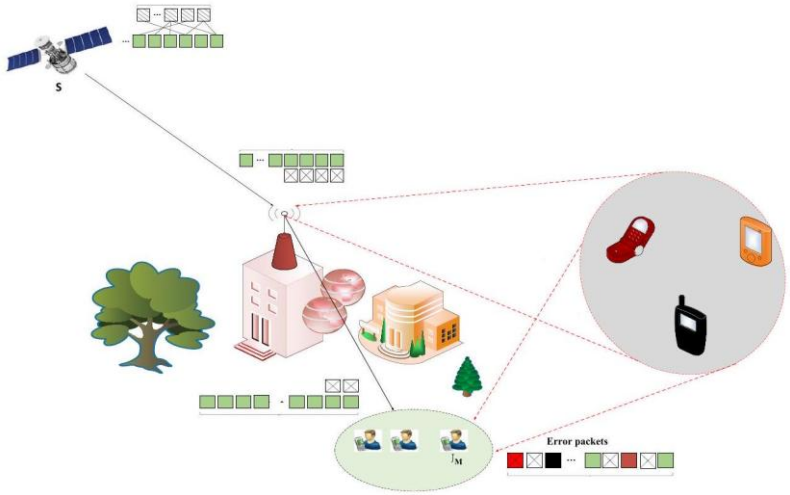


Figure 1. The proposed FC-aided HSTRs.

Let g_{AB} denote channel gain of the A → B channels, where A and B are a transmitter and a receiver, respectively, $(\mathrm{A},\mathrm{B}) \in \{\mathrm{S},\mathrm{R},\mathrm{U}_m,\mathrm{I}_k\}, m = 1,2,\ldots,M, k = 1,2,\ldots,K$. Assume that all A → B channels are block and flat fading, meaning that g_{AB} does not change during each transmission of p_{S}, but varies independently after each transmission of p_{S}.

For the S → R link, the channel gain g_{SR} which experiences a Shadowed-Rician distribution has the following PDF (see [30]-[31]):

$$f_{\gamma_{\mathrm{SR}}}(x) = \frac{1}{2b_{\mathrm{SR}}}\left(\frac{2a_{\mathrm{SR}}b_{\mathrm{SR}}}{2a_{\mathrm{SR}}b_{\mathrm{SR}}+\Omega_{\mathrm{SR}}}\right)^{a_{\mathrm{SR}}} \exp\left(-\frac{x}{2b_{\mathrm{SR}}}\right) {}_1F_1\left(a_{\mathrm{SR}};1;\frac{\Omega_{\mathrm{SR}}x}{2b_{\mathrm{SR}}(2a_{\mathrm{SR}}b_{\mathrm{SR}}+\Omega_{\mathrm{SR}})}\right), \quad (1)$$

where $2b_{\mathrm{SR}}$ and Ω_{SR} indicate the mean values of the multi-path and Line of Sight (LOS) components, respectively, a_{SR} is a fading parameter, and ${}_1F_1(.;.;.)$ is a confluent hypergeometric function of the first kind [30]-[31].

Using [34], CDF of g_{SR} can be expressed under the following form:

$$\begin{aligned} F_{\gamma_{\mathrm{SR}}}(x) &= 1 - \alpha_{\mathrm{SR}}^{a_{\mathrm{SR}}}\psi_{\mathrm{SR}} \sum_{n_{\mathrm{SR}}=0}^{a_{\mathrm{SR}}-1} \sum_{q_{\mathrm{SR}}=0}^{n_{\mathrm{SR}}} \frac{(n_{\mathrm{SR}})!}{(q_{\mathrm{SR}})!} \frac{\xi_{\mathrm{SR}}(n_{\mathrm{SR}})}{(\psi_{\mathrm{SR}}-\beta_{\mathrm{SR}})^{n_{\mathrm{SR}}-q_{\mathrm{SR}}+1}} x^{q_{\mathrm{SR}}} \exp(-(\psi_{\mathrm{SR}}-\beta_{\mathrm{SR}})x) \\ &= 1 - \sum_{n_{\mathrm{SR}}=0}^{a_{\mathrm{SR}}-1} \sum_{q_{\mathrm{SR}}=0}^{n_{\mathrm{SR}}} \Psi_0 x^{q_{\mathrm{SR}}} \exp(-(\psi_{\mathrm{SR}}-\beta_{\mathrm{SR}})x), \end{aligned} \quad (2)$$

where (n_{ST}) is Pochhammer function [34], and

$$\psi_{\mathrm{SR}} = \frac{1}{2b_{\mathrm{SR}}}, \alpha_{\mathrm{SR}} = \left(\frac{2a_{\mathrm{SR}}b_{\mathrm{SR}}}{2a_{\mathrm{SR}}b_{\mathrm{SR}}+\Omega_{\mathrm{SR}}}\right)^{a_{\mathrm{SR}}}, \beta_{\mathrm{SR}} = \left(\frac{\Omega_{\mathrm{SR}}}{2b_{\mathrm{SR}}(2a_{\mathrm{SR}}b_{\mathrm{SR}}+\Omega_{\mathrm{SR}})}\right),$$

$$\xi_{\mathrm{SR}}(n_{\mathrm{SR}}) = \frac{(-1)^{n_{\mathrm{SR}}}(1-a_{\mathrm{SR}})\beta_{\mathrm{SR}}^{n_{\mathrm{SR}}}}{(n_{\mathrm{SR}})!}, \Psi_0 = \frac{(n_{\mathrm{SR}})!}{(q_{\mathrm{SR}})!} \frac{\alpha_{\mathrm{SR}}^{a_{\mathrm{SR}}}\psi_{\mathrm{SR}}\xi_{\mathrm{SR}}(n_{\mathrm{SR}})}{(\psi_{\mathrm{SR}}-\beta_{\mathrm{SR}})^{n_{\mathrm{SR}}-q_{\mathrm{SR}}+1}}. \quad (3)$$

For the $\mathrm{R} \rightarrow \mathrm{U}_m, \mathrm{I}_k \rightarrow \mathrm{R}$ and $\mathrm{I}_k \rightarrow \mathrm{U}_m$ links, all these channels are assumed to be Rayleigh fading. Hence, $g_{\mathrm{RU}_m}, g_{\mathrm{I}_k\mathrm{R}}$ and $g_{\mathrm{I}_k\mathrm{U}_m}$ experience exponential distributions, and their PDFs can be written, respectively, as [36]:

$$f_{g_{\mathrm{RU}_m}}(x) = \lambda_{\mathrm{RU}_m}\exp\left(-\lambda_{\mathrm{RU}_m}x\right), f_{g_{\mathrm{t}\mathrm{R}_\mathrm{R}}}(x) = \lambda_{\mathrm{I}_k\mathrm{R}}\exp\left(-\lambda_{\mathrm{I}_k\mathrm{R}}x\right), f_{g_{\mathrm{I}_k\mathrm{U}_m}}(x) = \lambda_{\mathrm{I}_k\mathrm{U}_m}\exp\left(-\lambda_{\mathrm{I}_k\mathrm{U}_m}x\right), \quad (4)$$

where $\lambda_{\mathrm{RU}_m}, \lambda_{\mathrm{I}_k\mathrm{R}}$ and $\lambda_{\mathrm{I}_k\mathrm{R}_m}$ are fading parameters of the $\mathrm{R} \rightarrow \mathrm{U}_m, \mathrm{I}_k \rightarrow \mathrm{R}$ and $\mathrm{I}_k \rightarrow \mathrm{U}_m$ links, respectively [1].

As explained in [37], for ease of presentation and analytical tractability, the channel coefficients $g_{\mathrm{RU}_m}, g_{\mathrm{I}_k\mathrm{R}}$ and $g_{\mathrm{I}_k\mathrm{U}_m}$ can be assumed to be independent and identically distributed (i.i.d.). Similarly, all interference sources are assumed to have identical transmit power and experience i.i.d. Rayleigh fading, which enables the aggregate interference to be modelled in a tractable form [38]. Accordingly, the large-scale fading parameters satisfy $\lambda_{\mathrm{RU}_m} = \lambda_{\mathrm{RU}}(\forall m), \lambda_{\mathrm{I}_k\mathrm{R}} = \lambda_{\mathrm{IR}}(\forall k)$ and $\lambda_{\mathrm{I}_k\mathrm{U}_m} = \lambda_{\mathrm{IU}}(\forall k, m)$. This common assumption allows the decoding events to maintain a binomial structure and enables closed-form derivations in the subsequent analysis. Therefore, we can rewrite (4) under the following forms:

$$f_{g_{\mathrm{RU}_m}}(x) = \lambda_{\mathrm{RU}}\exp(-\lambda_{\mathrm{RU}}x), f_{g_{\mathrm{I}_\mathrm{R}\mathrm{R}}}(x) = \lambda_{\mathrm{IR}}\exp(-\lambda_{\mathrm{IR}}x), f_{g_{\mathrm{I}_k\mathrm{U}_m}}(x) = \lambda_{\mathrm{IU}}\exp(-\lambda_{\mathrm{IU}}x) \quad (5)$$

From (5), the corresponding CDFs can be obtained, respectively, as:

$$F_{g_{\mathrm{RU}_m}}(x) = 1 - \exp(-\lambda_{\mathrm{RU}}x), F_{g_{I_k\mathrm{R}}}(x) = 1 - \exp(-\lambda_{\mathrm{IR}}x), F_{g_{I_k\mathrm{U}_m}}(x) = 1 - \exp(-\lambda_{\mathrm{IU}}x) \quad (6)$$

Next, the operational principles of the conventional FC-aided HSTRNs (ConV) and the proposed FC-aided HSTRNs (ProP) are described in detail.

2.1 The ConV Scheme

In the ConV scheme, the relay station R forwards each encoded packet p_S to the ground users, without storing any p_S in its buffer. In particular, at the first time slot, S transmits p_S to R, and the instantaneous SNR obtained at R can be given as [39]:

$$\psi_{\mathrm{SR}}^{\mathrm{ConV}} = \frac{P_\mathrm{S} g_{\mathrm{SR}}}{\sum_{k=1}^{K} P_\mathrm{I} g_{\mathrm{I}_k\mathrm{R}} + \sigma_0^2} = \frac{\Delta_\mathrm{S} g_{\mathrm{SR}}}{\sum_{k=1}^{K} \Delta_\mathrm{I} g_{\mathrm{I}_k\mathrm{R}} + 1}, \quad (7)$$

where P_S and P_I are transmit power of the satellite S and all interference sources, respectively, σ_0^2 is variance of Gaussian noises at all receivers B, $\Delta_\mathrm{S} = P_\mathrm{S}/\sigma_0^2$ and $\Delta_\mathrm{I} = P_\mathrm{I}/\sigma_0^2$ denotes transmit SNRs.

If R decodes p_S successfully, it sends p_S to all the ground users at the second time slot, using the DF approach. The instantaneous SNR of the $\mathrm{R} \rightarrow \mathrm{U}_m$ link can be given as:

$$\psi_{\mathrm{RU}_m}^{\mathrm{ConV}} = \frac{P_\mathrm{R} g_{\mathrm{RU}_m}}{\sum_{k=1}^{K} P_\mathrm{I} g_{\mathrm{I}_k\mathrm{U}_m} + \sigma_0^2} = \frac{\Delta_\mathrm{R} g_{\mathrm{RU}_m}}{\sum_{k=1}^{K} \Delta_\mathrm{I} g_{\mathrm{I}_k\mathrm{U}_m} + 1}, \quad (8)$$

where P_R is transmit power of R , and $\Delta_\mathrm{R} = P_\mathrm{R}/\sigma_0^2$.

We now consider the time allocation between the first and second time slots in the ConV scheme. Assume that the total duration for the $\mathrm{S} \rightarrow \mathrm{R} \rightarrow \mathrm{U}$ transmission is 01-time unit. Let the time allocated to the first and second time slots be τ_{ConV} and $1 - \tau_{\mathrm{ConV}}$, respectively, where $0 < \tau_{\mathrm{ConV}} < 1$. Therefore, we can formulate the instantaneous channel capacity of the $\mathrm{S} \rightarrow \mathrm{R}$ and $\mathrm{R} \rightarrow \mathrm{U}_m$ links, respectively, as:

$$C_{\mathrm{SR}}^{\mathrm{ConV}} = \tau_{\mathrm{ConV}} \log_2\left(1 + \psi_{\mathrm{SR}}^{\mathrm{ConV}}\right),$$

$$C_{\mathrm{RU}_m}^{\mathrm{ConV}} = (1 - \tau_{\mathrm{ConV}}) \log_2\left(1 + \psi_{RU_m}^{ConV}\right) \quad (9)$$

2.2 The ProP Scheme

Similar to the ConV scheme, S sends p_S to R at the first time slot, and the instantaneous SNR obtained at R can be given, similarly, as (7):

$$\psi_{\mathrm{SR}}^{\mathrm{ProP}} = \frac{\Delta_\mathrm{S} g_{\mathrm{SR}}}{\sum_{k=1}^{K} \Delta_\mathrm{I} g_{\mathrm{I}_k\mathrm{R}} + 1}, \quad (10)$$

Let τ_{ProP} and $1-\tau_{\text{ProP}}$ denote the time allocated to the first and second time slots in the ProP scheme, where $0<\tau_{\text{ProP}}<1$. Then, the instantaneous channel capacity of the S → R link can be given as:

$$C_{\text{SR}}^{\text{PrOP}}=\tau_{\text{ProP}}\log_2\left(1+\psi_{\text{SR}}^{\text{PrOP}}\right) \tag{11}$$

If the decoding at R is successful, R sends p_{S} to all the ground users at the second time slot. Then, the instantaneous SNR of the R → U_m link can be given, similarly, as (8):

$$\psi_{\text{RU}_m}^{\text{ProP}^{\text{Pr}}}=\frac{\Delta_{\text{R}}g_{\text{RU}_m}}{\sum_{k=1}^{K}\Delta_{\text{I}}g_{\text{I}_k\text{U}_m}+1}. \tag{12}$$

Then, the instantaneous channel capacity of the R → U_m link can be expressed as:

$$C_{\text{RU}_m}^{\text{ProP,Case 1}}=(1-\tau_{\text{ProP}})\log_2\left(1+\psi_{\text{RU}_m}^{\text{ProP}}\right). \tag{13}$$

As mentioned, the relay station in ProP stores in its buffers the encoded packets which are correctly decoded. When the number of packets p_{S} correctly obtained at R equals $G_{\min}$, it can recover the original data. Moreover, R will replace S to transmit encoded packets to the ground users. After this point, the relay switches to the transmission phase and forwards encoded packets to the ground users. Due to the half-duplex constraint, the relay no longer attempts to receive additional packets from the satellite once it starts forwarding. Meanwhile, the satellite may continue transmitting up to $N_{\max}$, as limited by the system delay constraint. In addition, the packets transmitted by the relay are generated based on the recovered source data using Fountain coding, allowing the users to accumulate sufficient packets for successful decoding. In this case, the instantaneous channel capacity of the R → U_m link can be calculated as:

$$C_{\text{RU}_m}^{\text{ProP,Case 2}}=\log_2\left(1+\psi_{\text{RU}_m}^{\text{ProP}}\right) \tag{14}$$

Remark 1: The absence of a direct satellite-user link in the considered model represents a worst-case scenario commonly adopted in hybrid satellite-terrestrial networks. If a weak direct link is present, its contribution can be incorporated as an additional SINR component without altering the structure of the proposed ProP scheme. In this case, both OP and SOP are expected to improve due to additional diversity, while the relative performance gain of ProP over the conventional scheme remains intact.

3. PERFORMANCE ANALYSIS

This section derives exact closed-form expressions of OP and SOP for the ConV and ProP schemes. At first, we evaluate the decoding probability of one encoded packet p_{S}.

3.1 Decoding of the Packet p_{S}

It is worth noting that one packet p_{S} is successfully transmitted to U_m if both S → R and R → U_m transmissions are successful.

Indeed, considering the transmission of p_{s} at the first time slot in the ConV scheme; from (7) and (9), the probability of the correct decoding of p_{S} at R can be formulated as:

$$\begin{aligned}\theta_{\text{SR}}^{\text{ConV}}&=\Pr\left(C_{\text{SR}}^{\text{ConV}}\geq C_{1,\text{th}}\right)=\Pr\left(\psi_{\text{SR}}^{\text{ConV}}\geq\psi_{1,\text{th}}^{\text{ConV}}\right)=\Pr\left(g_{\text{SR}}\geq\frac{\Delta_{\text{I}}\psi_{1,\text{th}}^{\text{ConV}}}{\Delta_{\text{S}}}\sum_{k=1}^{K}g_{\text{I}_k\text{R}}+\frac{\psi_{1,\text{th}}^{\text{ConV}}}{\Delta_{\text{S}}}\right)\\&=\Pr\left(g_{\text{SR}}\geq\rho_{1,\text{th}}^{\text{ConV}}Z_{\text{R,Sum}}+\rho_{2,\text{th}}^{\text{ConV}}\right)\end{aligned} \tag{15}$$

where $C_{1,\text{ th}}$ is a target rate of the first link between S and R, $Z_{\text{R, Sum}}=\sum_{k=1}^{K}g_{\text{I}_k\text{R}}$, and

$$\psi_{1,\text{th}}^{\text{ConV}}=2^{\frac{C_{1,\text{th}}}{\tau_{\text{ConV}}}}-1,\rho_{1,\text{th}}^{\text{ConV}}=\frac{\Delta_{\text{I}}\psi_{1,\text{th}}^{\text{ConV}}}{\Delta_{\text{S}}},\rho_{2,\text{th}}^{\text{ConV}}=\frac{\psi_{1,\text{th}}^{\text{ConV}}}{\Delta_{\text{S}}} \tag{16}$$

Since $Z_{\text{R, Sum}}$ is a summation of K exponential random variables, using [40, eq. (B.9)], we obtain PDF of $Z_{\text{R,Sum}}$ as:

$$f_{Z_{\text{R, Sum}}}(x)=\frac{(\lambda_{\text{IR}})^K x^{K-1}\exp(-\lambda_{\text{IR}}x)}{(K-1)!} \tag{17}$$

Next, we can express $\theta_{\mathrm{SR}}^{\mathrm{ConV}}$ in (15) under the following form:

$$\theta_{\mathrm{SR}}^{\mathrm{ConV}} = \int_0^{+\infty} \left(1 - F_{g_{\mathrm{SR}}}\left(\rho_{1,\mathrm{th}}^{\mathrm{ConV}} x + \rho_{2,\mathrm{th}}^{\mathrm{ConV}}\right)\right) f_{Z_{\mathrm{R,Sum}}}(x) dx \tag{18}$$

From (2), we can express $F_{g_{\mathrm{SR}}}\left(\rho_{1,\mathrm{th}}^{\mathrm{ConV}} x + \rho_{2,\mathrm{th}}^{\mathrm{ConV}}\right)$ under the following form:

$$\begin{aligned} F_{\gamma_{\mathrm{SR}}}\left(\rho_{1,\mathrm{th}}^{\mathrm{ConV}} x + \rho_{2,\mathrm{th}}^{\mathrm{ConV}}\right) &= 1 - \sum_{n_{\mathrm{SR}}=0}^{a_{\mathrm{SR}}-1} \sum_{q_{\mathrm{SR}}=0}^{n_{\mathrm{SR}}} \Psi_1^{\mathrm{ConV}} \left(x + \rho_{3,\mathrm{th}}^{\mathrm{ConV}}\right)^{q_{\mathrm{SR}}} \exp\left(-\rho_{4,\mathrm{th}}^{\mathrm{ConV}} x\right) \\ &= 1 - \sum_{n_{\mathrm{SR}}=0}^{a_{\mathrm{SR}}-1} \sum_{q_{\mathrm{SR}}=0}^{n_{\mathrm{SR}}} \sum_{t_{\mathrm{SR}}=0}^{q_{\mathrm{SR}}} \Psi_1^{\mathrm{ConV}} \binom{q_{\mathrm{SR}}}{t_{\mathrm{SR}}} \left(\rho_{3,\mathrm{th}}^{\mathrm{ConV}}\right)^{t_{\mathrm{SR}}} x^{q_{\mathrm{SR}}-t_{\mathrm{SR}}} \exp\left(-\rho_{4,\mathrm{th}}^{\mathrm{ConV}} x\right) \\ &= 1 - \sum_{n_{\mathrm{SR}}=0}^{a_{\mathrm{SR}}-1} \sum_{q_{\mathrm{SR}}=0}^{n_{\mathrm{SR}}} \sum_{t_{\mathrm{SR}}=0}^{q_{\mathrm{SR}}} \Psi_2^{\mathrm{ConV}} x^{q_{\mathrm{SR}}-t_{\mathrm{SR}}} \exp\left(-\rho_{4,\mathrm{th}}^{\mathrm{ConV}} x\right), \end{aligned} \tag{19}$$

where

$$\rho_{3,\mathrm{th}}^{\mathrm{ConV}} = \frac{\rho_{2,\mathrm{th}}^{\mathrm{ConV}}}{\rho_{1,\mathrm{th}}^{\mathrm{ConV}}}, \rho_{4,\mathrm{th}}^{\mathrm{ConV}} = (\psi_{\mathrm{SR}} - \beta_{\mathrm{SR}}) \rho_{1,\mathrm{th}}^{\mathrm{ConV}}, \binom{q_{\mathrm{SR}}}{t_{\mathrm{SR}}} = \frac{(q_{\mathrm{SR}})!}{(t_{\mathrm{SR}})!\,(q_{\mathrm{SR}} - t_{\mathrm{SR}})!}$$

$$\Psi_1^{\mathrm{ConV}} = \Psi_0 \left(\rho_{1,\mathrm{th}}^{\mathrm{ConV}}\right)^{q_{\mathrm{SR}}} \exp\left(-(\psi_{\mathrm{SR}} - \beta_{\mathrm{SR}}) \rho_{2,\mathrm{th}}^{\mathrm{ConV}}\right), \Psi_2^{\mathrm{ConV}} = \binom{q_{\mathrm{SR}}}{t_{\mathrm{SR}}} \left(\rho_{3,\mathrm{th}}^{\mathrm{ConV}}\right)^{t_{\mathrm{SR}}} \Psi_1. \tag{20}$$

Substituting (17) and (19) into (18), after some manipulations, we obtain an exact closed-form expression of $\theta_{\mathrm{SR}}^{\mathrm{ConV}}$ as:

$$\begin{aligned} \theta_{\mathrm{SR}}^{\mathrm{ConV}} &= \sum_{n_{\mathrm{SR}}=0}^{a_{\mathrm{SR}}-1} \sum_{q_{\mathrm{SR}}=0}^{n_{\mathrm{SR}}} \sum_{t_{\mathrm{SR}}=0}^{q_{\mathrm{SR}}} \frac{(\lambda_{\mathrm{IR}})^K \Psi_2}{(K-1)!} \int_0^{+\infty} x^{K+q_{\mathrm{SR}}-t_{\mathrm{SR}}-1} \exp\left(-\left(\rho_{4,\mathrm{th}}^{\mathrm{ConV}} + \lambda_{\mathrm{IR}}\right) x\right) dx \\ &= \sum_{n_{\mathrm{SR}}=0}^{a_{\mathrm{SR}}-1} \sum_{q_{\mathrm{SR}}=0}^{n_{\mathrm{SR}}} \sum_{t_{\mathrm{SR}}=0}^{q_{\mathrm{SR}}} \frac{(K + q_{\mathrm{SR}} - t_{\mathrm{SR}} - 1)!}{(K-1)!} \frac{(\lambda_{\mathrm{IR}})^K \Psi_2^{\mathrm{ConV}}}{\left(\lambda_{\mathrm{IR}} + \rho_{4,\mathrm{th}}^{\mathrm{ConV}}\right)^{K+q_{\mathrm{SR}}-t_{\mathrm{SR}}}} \end{aligned} \tag{21}$$

Considering the transmission at the second time slot in the ConV scheme; the probability of the successful decoding of one encoded packet p_{S} at U_m can be formulated by using (8) and (9) as:

$$\begin{aligned} \theta_{\mathrm{RU}_m}^{\mathrm{ConV}} = \Pr\left(C_{\mathrm{RU}_m}^{\mathrm{ConV}} \geq C_{2,\mathrm{th}}\right) = \Pr\left(\psi_{\mathrm{RU}_m}^{\mathrm{ConV}} \geq \psi_{2,\mathrm{th}}^{\mathrm{ConV}}\right) = \Pr\left(g_{\mathrm{RU}_m} \geq \frac{\Delta_{\mathrm{I}} \psi_{2,\mathrm{th}}^{\mathrm{ConV}}}{\Delta_{\mathrm{R}}} \sum_{k=1}^{K} g_{\mathrm{I}_k \mathrm{U}_m} + \frac{\psi_{2,\mathrm{th}}^{\mathrm{ConV}}}{\Delta_{\mathrm{R}}}\right) \\ = \Pr\left(g_{\mathrm{RU}_m} \geq \omega_{1,\mathrm{th}}^{\mathrm{ConV}} Z_{\mathrm{U}_m,\mathrm{Sum}} + \omega_{2,\mathrm{th}}^{\mathrm{ConV}}\right) \end{aligned} \tag{22}$$

where $C_{2,\mathrm{th}}$ is a target rate of the $\mathrm{R} \rightarrow \mathrm{U}_m$ link, $Z_{\mathrm{U}_m,\mathrm{Sum}} = \sum_{k=1}^{K} g_{\mathrm{I}_k \mathrm{U}_m}$, and

$$\psi_{2,\mathrm{th}}^{\mathrm{ConV}} = 2^{\frac{C_{2,\mathrm{th}}}{1-\tau_{\mathrm{ConV}}}} - 1, \omega_{1,\mathrm{th}}^{\mathrm{ConV}} = \frac{\Delta_{\mathrm{I}} \psi_{2,\mathrm{th}}^{\mathrm{ConV}}}{\Delta_{\mathrm{R}}}, \omega_{2,\mathrm{th}}^{\mathrm{ConV}} = \frac{\psi_{2,\mathrm{th}}^{\mathrm{ConV}}}{\Delta_{\mathrm{R}}}. \tag{23}$$

Similar to (17), PDF of $Z_{\mathrm{U}_m,\,\mathrm{Sum}}$ can be expressed as:

$$f_{Z_{\mathrm{U}_m,\,\mathrm{Sum}}}(x) = \frac{(\lambda_{\mathrm{IU}})^M x^{M-1} \exp(-\lambda_{\mathrm{IU}} x)}{(M-1)!} \tag{24}$$

Substituting (6) and (24) into (22), after some manipulations, we obtain an exact closed-form expression of $\theta_{\mathrm{RU}_m}^{\mathrm{ConV}}$ as:

$$\begin{aligned} \theta_{\mathrm{RU}_m}^{\mathrm{ConV}} &= \int_0^{+\infty} \left[1 - F_{g_{\mathrm{RU}_m}}\left(\omega_{1,\mathrm{th}}^{\mathrm{ConV}} x + \omega_{2,\mathrm{th}}^{\mathrm{ConV}}\right)\right] f_{Z_{\mathrm{U}_m,\mathrm{Sum}}}(x) dx \\ &= \frac{(\lambda_{\mathrm{IU}})^M}{(M-1)!} \exp\left(-\lambda_{\mathrm{RU}} \omega_{2,\mathrm{th}}^{\mathrm{ConV}}\right) \int_0^{+\infty} x^{M-1} \exp\left(-\left(\lambda_{\mathrm{IU}} + \lambda_{\mathrm{RU}} \omega_{1,\mathrm{th}}^{\mathrm{ConV}}\right) x\right) dx \\ &= \left(\frac{\lambda_{\mathrm{IU}}}{\lambda_{\mathrm{IU}} + \lambda_{\mathrm{RU}} \omega_{1,\mathrm{th}}^{\mathrm{ConV}}}\right)^M \exp\left(-\lambda_{RU} \omega_{2,th}^{ConV}\right) \end{aligned} \tag{25}$$

Then, the probability of the successful decoding of one packet p_S at U_m in the ConV scheme is computed as:

$$\theta_{U_m}^{\text{ConV}} = \theta_{SR}^{\text{ConV}} \theta_{RU_m}^{\text{ConV}} \tag{26}$$

Considering the ProP scheme; the probability that one packet p_S is decoded correctly by the user U_m can be calculated in two cases as follows: Case 1: p_S is sent to U_m via the S → R → U_m link; Case 2: p_S is directly transmitted from R to U_m.

Case 1: p_S is sent to U_m *via* the S → R → U_m link

Similar to (21) and (24), the probability that one packet p_S is decoded correctly at the S → R and R → U_m links can be given, respectively, as:

$$\theta_{SR}^{\text{ProP}} = \sum_{n_{SR}=0}^{a_{SR}-1} \sum_{q_{SR}=0}^{n_{SR}} \sum_{t_{SR}=0}^{q_{SR}} \frac{(K+q_{SR}-t_{SR}-1)!}{(K-1)!} \frac{(\lambda_{IR})^K \Psi_2^{\text{ProP}}}{\left(\lambda_{IR}+\rho_{4,\text{th}}^{\text{ProP}}\right)^{K+q_{SR}-t_{SR}}}, \tag{27}$$

$$\theta_{RU}^{\text{ProP, Case 1}} = \left(\frac{\lambda_{IU}}{\lambda_{IU}+\lambda_{RU}\omega_{1,\text{th}}^{\text{ProP}}}\right)^M \exp\left(-\lambda_{RU}\omega_{2,\text{th}}^{\text{ProP}}\right), \tag{28}$$

where

$$\psi_{1,\text{th}}^{\text{ProP}} = 2^{\frac{C_{1,\text{th}}}{\tau_{\text{ProP}}}} - 1, \rho_{1,\text{th}}^{\text{ProP}} = \frac{\Delta_I \psi_{1,\text{th}}^{\text{ProP}}}{\Delta_S}, \rho_{2,\text{th}}^{\text{ProP}} = \frac{\psi_{1,\text{th}}^{\text{ProP}}}{\Delta_S}, \rho_{3,\text{th}}^{\text{ProP}} = \frac{\rho_{2,\text{th}}^{\text{ProP}}}{\rho_{1,\text{th}}^{\text{ProP}}}, \rho_{4,\text{th}}^{\text{ProP}} = (\psi_{SR}-\beta_{SR})\rho_{1,\text{th}}^{\text{ProP}},$$

$$\Psi_1^{\text{ProP}} = \Psi_0 \left(\rho_{1,\text{th}}^{\text{ProP}}\right)^{q_{SR}} \exp\left(-(\psi_{SR}-\beta_{SR})\rho_{2,\text{th}}^{\text{ProP}}\right), \Psi_2^{\text{ProP}} = \binom{q_{SR}}{t_{SR}} \left(\rho_{3,\text{th}}^{\text{ProP}}\right)^{t_{SR}} \Psi_1^{\text{ProP}},$$

$$\psi_{2,\text{th}}^{\text{ProP}} = 2^{\frac{C_{2,\text{th}}}{1-\tau_{\text{ConV}}}} - 1, \omega_{1,\text{th}}^{\text{ProP}} = \frac{\Delta_I \psi_{2,\text{th}}^{\text{ProP}}}{\Delta_R}, \omega_{2,\text{th}}^{\text{ProP}} = \frac{\psi_{2,\text{th}}^{\text{ProP}}}{\Delta_R}, \tag{29}$$

Case 2: p_S is directly transmitted from R to U_m

In this case, using (14) instead of (13) for Case 2, we can obtain the probability that p_S is successfully transmitted from R to U_m as

$$\theta_{RU_m}^{\text{ProP,Case2}} = \left(\frac{\lambda_{IU}}{\lambda_{IU}+\lambda_{RU}\omega_{3,\text{th}}^{\text{ProP}}}\right)^M \exp\left(-\lambda_{RU}\omega_{4,\text{th}}^{\text{ProP}}\right), \tag{30}$$

where

$$\psi_{3,\text{th}}^{\text{PrOP}} = 2^{C_{2,\text{th}}} - 1, \omega_{3,\text{th}}^{\text{ProP}} = \frac{\Delta_I \psi_{3,\text{th}}^{\text{ProP}}}{\Delta_R}, \omega_{4,\text{th}}^{\text{PrOP}} = \frac{\psi_{3,\text{th}}^{\text{ProP}}}{\Delta_R} \tag{31}$$

3.2 Outage Probability (OP) at Each User

Considering the ConV scheme; let denote L_m^{ConV} as the number of p_S that is correctly received at U_m. If $L_m^{\text{ConV}} < G_{\min}$, U_m cannot reconstruct the original data of the satellite. Hence, OP at U_m in ConV can be expressed by an exact closed-form expression as:

$$\begin{aligned} \text{OP}_m^{\text{ConV}} &= 1 - \sum_{T_S=G_{\min}}^{N_{\max}} \Pr\left(L_R^{\text{ConV}} = G_{\min}, T_S\right) \Pr\left(L_m^{\text{ConV}} \geq G_{\min} \mid T_S\right) \\ &= \sum_{L_m^{\text{ConV}}=0}^{G_{\min}-1} \binom{N_{\max}}{L_m^{\text{ConV}}} \left(\theta_{U_m}^{\text{ConV}}\right)^{L_m^{\text{ConV}}} \left(1-\theta_{U_m}^{\text{ConV}}\right)^{N_{\max}-L_m^{\text{ConV}}} \end{aligned} \tag{32}$$

where $\left(1-\theta_{U_m}^{\text{ConV}}\right)$ is the probability that p_S cannot be successfully reached to U_m. For the ProP scheme; we consider the probability that the user U_m can successfully recover the original data of the satellite. Indeed, if we denote L_m^{ProP} as the number of p_S that is correctly received at U_m, then $L_m^{\text{ProP}} \geq G_{\min}$. It is worth noting that if L_R^{ProP} denotes the number of p_S being correctly received at R, then L_R^{ProP} must be equal to $G_{\min}$. It is because if $L_R^{\text{ProP}} < G_{\min}$, then $L_m^{\text{ProP}} < L_R^{\text{ProP}} < G_{\min}$, and U_m is

then in outage. Then, let us denote T_{S} as the number of transmission times of the satellite until the station R collects enough $G_{\min}$ encoded packets p_{S}, where $G_{\min} \le T_{\mathrm{S}} \le N_{\max}$. Therefore, the number of transmission times of R in Case 2 is given as $T_{\mathrm{R}} = N_{\max} - T_{\mathrm{S}}$. We also denote $L_{m,\,\mathrm{Casel}}^{\mathrm{ProP}}$ as the number of p_{S} that is correctly received at U_m in Case 1, where $0 \le L_{m,\,\mathrm{Casel}}^{\mathrm{ProP}} \le T_{\mathrm{S}}$. Finally, the successful reconstruction of the original data at U_m in ProP can be formulated as:

$$\overline{\mathrm{OP}}_m^{\mathrm{ProP}} = 1 - \sum_{T_{\mathrm{S}}=G_{\min}}^{N_{\max}} \Pr\left(L_R^{\mathrm{ProP}} = G_{\min}, T_{\mathrm{S}}\right) \Pr\left(L_{m,\,\mathrm{Case1}}^{\mathrm{ProP}} + L_{m,\,\mathrm{Case\,2}}^{\mathrm{ProP}} \ge G_{\min} \mid T_{\mathrm{S}}\right). \quad (33)$$

In (33), $\Pr\left(L_{\mathrm{R}}^{\mathrm{ProP}} = G_{\min}, T_{\mathrm{S}}\right) = \binom{T_{\mathrm{S}}-1}{G_{\min}-1} \left(\theta_{\mathrm{SR}}^{\mathrm{ProP}}\right)^{G_{\min}} \left(1-\theta_{\mathrm{SR}}^{\mathrm{ProP}}\right)^{T_{\mathrm{S}}-G_{\min}}$ is the probability that R collects enough $G_{\min}$ encoded packets p_{S} after T_{S} transmission times of S. It is worth noting that the transmission between S and R at the T_{S}-th transmission must be successful, and R must correctly receive $G_{\min} - 1$ packets p_{S} previously.

$\binom{G_{\min}}{L_{m,\,\mathrm{Case\,1}}^{\mathrm{ProP}}} \left(\theta_{\mathrm{RU}_m}^{\mathrm{ProP,Case\,1}}\right)^{L_{m,\,\mathrm{Casel}}^{\mathrm{Prop}}} \left(1-\theta_{\mathrm{RU}_m}^{\mathrm{ProP,Case\,1}}\right)^{G_{\min}-L_{m,\,\mathrm{Casel}}^{\mathrm{ProP}}}$ is the probability that U_m correctly receives $L_{m,\,\mathrm{Casel}}^{\mathrm{ProP}}$ packets p_{S} in Case 1, and $\mathrm{OP}_m^{\mathrm{ProP}} \binom{N_{\max}-T_{\mathrm{S}}}{L_{m,\,\mathrm{Case1}}^{\mathrm{ProP}}} \left(\theta_{\mathrm{RU}_m}^{\mathrm{ProP,Case\,2}}\right)^{L_{m,\,\mathrm{Case2}}^{\mathrm{ProP}}} \left(1-\theta_{\mathrm{RU}_m}^{\mathrm{ProP,Case\,2}}\right)^{N_{\max}-T_{\mathrm{S}}-L_{m,\,\mathrm{Case\,2}}^{\mathrm{Prop}}}$ is the probability that U_m correctly receives $L_{m,\,\mathrm{Case2}}^{\mathrm{ProP}}$ packets p_{S} in Case 2. Then, OP at U_m in ProP is obtained as:

$$\mathrm{OP}_m^{\mathrm{ProP}} = 1 - \overline{\mathrm{OP}}_m^{\mathrm{ProP}} \quad (34)$$

Remark 2: Since the R → U and I → U links are independent and identical, it is straightforward that OP at all the users in the considered schemes is the same, i.e., we can write $\mathrm{OP}_m^{\mathrm{ConV}} = \mathrm{OP}^{\mathrm{ConV}}$ and $\mathrm{OP}_m^{\mathrm{ProP}} = \mathrm{OP}^{\mathrm{ProP}}$ for $\forall m$.

3.3 System Outage Probability (SOP)

Firstly, $\mathrm{SOP}_{\mathrm{X}}$ is defined as the probability that one of the users in the X scheme experiences an outage, where $\mathrm{X} \in \{\mathrm{ConV}, \mathrm{ProP}\}$. To obtain $\mathrm{SOP}_{\mathrm{X}}$, we have to consider the probability that all the users can successfully recover the original data of the satellite, i.e., $\overline{\mathrm{SOP}_{\mathrm{X}}} = 1 - \mathrm{SOP}_{\mathrm{X}}$.

Considering the ConV scheme; let us denote $L_{\mathrm{R}}^{\mathrm{ConV}}$ as the number of p_{S} that is correctly received by the station R after the transmission ends. In order that all the users can collect at least $G_{\min}$ packets p_{S}, we have $G_{\min} \le L_{\mathrm{R}}^{\mathrm{ConV}} \le N_{\max}$. Then, $\overline{\mathrm{SOP}}_{\mathrm{ConV}}$ can be computed as:

$$\begin{aligned}\overline{\mathrm{SOP}}_{\mathrm{ConV}} &= 1 - \sum_{L_{\mathrm{R}}^{\mathrm{ConV}}=G_{\min}}^{N_{\max}} \Pr\left(L_{\mathrm{R}}^{\mathrm{ConV}} = L_R\right) \Pr\left(L_m^{\mathrm{ConV}} \ge G_{\min}, \forall m \mid L_{\mathrm{R}}^{\mathrm{ConV}} = L_R\right) \\ &= \sum_{L_{\mathrm{R}}^{\mathrm{ConV}}=G_{\min}}^{N_{\max}} \left\{ \begin{array}{l} \binom{N_{\max}}{L_{\mathrm{R}}^{\mathrm{ConV}}} \left(\theta_{\mathrm{SR}}^{\mathrm{ConV}}\right)^{L_{\mathrm{R}}^{\mathrm{ConV}}} \left(1-\theta_{\mathrm{SR}}^{\mathrm{ConV}}\right)^{N_{\max}-L_{\mathrm{R}}^{\mathrm{ConV}}} \\ \times \prod_{m=1}^{M} \left[\sum_{L_m^{\mathrm{ConV}}=G_{\min}}^{L_{\mathrm{R}}^{\mathrm{ConV}}} \binom{L_m^{\mathrm{ConV}}}{L_{\mathrm{R}}^{\mathrm{ConV}}} \left(\theta_{\mathrm{RU}}^{\mathrm{ConV}}\right)^{L_m^{\mathrm{ConV}}} \left(1-\theta_{\mathrm{RU}}^{\mathrm{ConV}}\right)^{L_{\mathrm{R}}^{\mathrm{ConV}}-L_m^{\mathrm{ConV}}} \right] \end{array} \right\} \end{aligned} \quad (35)$$

In (35), $\Pr\left(L_m^{\mathrm{ConV}} \ge G_{\min}, \forall m \mid L_{\mathrm{R}}^{\mathrm{ConV}} = L_R\right) = \sum_{L_m^{\mathrm{ConV}}=G_{\min}}^{L_{\mathrm{R}}^{\mathrm{ConV}}} \binom{L_m^{\mathrm{ConV}}}{L_{\mathrm{R}}^{\mathrm{ConV}}} \left(\theta_{\mathrm{RU}_m}^{\mathrm{ConV}}\right)^{L_m^{\mathrm{ConV}}} \left(1-\theta_{\mathrm{RU}_m}^{\mathrm{ConV}}\right)^{L_{\mathrm{R}}^{\mathrm{ConV}}} - L_m^{\mathrm{ConV}}$ is the probability that U_m can successfully receive at least $G_{\min}$ encoded packets.

Considering the ProP scheme; using (33), we can obtain $\overline{\mathrm{SOP}}_{\mathrm{ProP}}$ as:

$$\overline{\mathrm{SOP}}_{\mathrm{ProP}} = 1 - \sum_{T_{\mathrm{S}}=G_{\min}}^{N_{\max}} \Pr\left(L_{\mathrm{R}}^{\mathrm{ProP}} = G_{\min}, T_S\right) \prod_{m=1}^{M} \Pr\left(L_{m,\,\mathrm{Case1}}^{\mathrm{ProP}} + L_{m,\,\mathrm{Case\,2}}^{\mathrm{ProP}} \ge G_{\min}, \forall m \mid T_S\right) \quad (36)$$

Finally, SOP of the ConV and ProP schemes can be obtained, respectively, as:

$$\mathrm{SOP}_{\mathrm{ConV}} = 1 - \overline{\mathrm{SOP}}_{\mathrm{ConV}}, \mathrm{SOP}_{\mathrm{ProP}} = 1 - \overline{\mathrm{SOP}}_{\mathrm{ProP}}. \quad (37)$$

Remark 3: It is worth emphasizing that the proposed FC-assisted HSTR framework fundamentally differs from non-FC schemes. While non-FC approaches typically define OP/SOP based on instantaneous SINR constraints at the symbol level, the FC-assisted scheme evaluates OP/SOP through packet-level decoding conditions governed by $G_{\min}$ and $N_{\max}$. This packet accumulation mechanism allows successful decoding as long as a sufficient number of packets is collected, thereby offering improved robustness against channel fading and co-channel interference.

3.4 Joint Time and Power Allocation Problem

Assume that the total transmit power of the S and R nodes is fixed such that $P_S + P_R = P_{tot}$. Specifically, we examine the following power-allocation scheme: $P_S = \mu_x P_{tot}, P_R = (1-\mu_x)P_{tot}$, where $X \in \{ConV, ProP\}$ and $0 < \mu_X < 1$. Now, we consider the joint time and power allocation problem as

$$\mathrm{Min}_{0<\tau_x<1, 0<\mu_x<1} \mathrm{SOP}_X. \quad (38)$$

Regarding the convergence of the adopted Golden-Section Search (GSS) algorithm, it is well known that GSS is guaranteed to converge for any continuous unimodal (or quasi-unimodal) function over a compact interval. In our optimization problem, the SOP-based objective function in (38) is continuous with respect to μ_X and τ_X on the bounded domain (0,1), which satisfies the classical convergence conditions of GSS [41]. In addition, we have numerically verified in all simulation settings that the iterative values μ_X^n and τ_X^n converge monotonically toward their optimal solutions within a small number of iterations (typically fewer than 20). This numerical behavior is fully consistent with the theoretical contraction factor $\varphi^{-1} \approx 0.618$ of GSS.

Moreover, the number of iterations is mainly determined by the stopping tolerance δ (commonly, $\delta = 10^{-4}: 10^{-3}$ is selected), and the GSS procedure is not sensitive to the choice of initial values, since the search is performed over predefined feasible intervals. Therefore, a convergence figure is not required, and the algorithm's convergence is theoretically ensured and empirically confirmed [42]. It is worth noting that the proposed optimization operates over low-dimensional variables and does not scale combinatorially with the number of users. The computational complexity of the GSS-based method is logarithmic with respect to the desired accuracy, which ensures efficient convergence. Therefore, the proposed approach remains computationally efficient and practically feasible even in large-scale user scenarios.

4. RESULTS

This section presents Monte-Carlo simulations (Sim) to verify the exact closed-form expressions (Theory) of OP and SOP for the ConV and ProP schemes. The simulation parameters are selected based on commonly adopted settings in the literature [12], [14], [31], [34]. Unless otherwise stated, the simulation parameters are set as follows: the Shadowed-Rician parameters $(a_{SR}, b_{SR}, \Omega_{SR})$ are (1,0.063,0.0007) or FHS and (5,0.251,0.279) for AS; the target rates are $C_{1,\mathrm{th}} = 0.2$, and $C_{2,\mathrm{th}} = 0.1$ and the average channel gains are $\lambda_{RU} = 1, \lambda_{IR} = 50$ and $\lambda_{IU} = 25$. In addition, $K = 3, G_{\min} = 6, \sigma_0^2 = 1$, and $P_I = 0.25P_{tot}$, while the transmit SNR is defined as $\Delta = P_{tot}/\sigma_0^2$. In Figure 2, we additionally include one example under the FHS condition with $N_{\max} = 8$ to illustrate the performance degradation caused by severe shadowing. For the remaining figures (Figs. 3-8), we present results only under the AS channel, since both AS and FHS exhibit the same OP and SOP variation trends, and including all FHS cases would not provide further analytical insight. Prior studies, such as [21], [23], have also confirmed that FHS yields consistently worse performance than AS due to harsher propagation conditions.

In Figure 2, we present OP at each user in the ConV and ProP schemes as a function of transmit SNR($\Delta = P_{tot}/\sigma_0^2$) in dB when $\mu_{ConV} = \mu_{ProP} = 0.5$ and $\tau_{ConV} = \tau_{ProP} = 0.5$. At first, it is remined that OP at each user in the two considered schemes is the same. Next, we can observe that the OP performance of both ConV and ProP is better as Δ increases, since transmit power of the satellite and the relay station increases. However, as $\Delta \to +\infty$, it can be seen that OP of both ConV and ProP reaches saturation values. It is because at high Δ regions, SINRs of the S → R and R → U_m links do not depend on Δ. Indeed, with $P_S = \mu P_{tot}, P_R = (1-\mu)P_{tot}$ and $P_I = 0.25P_{tot}$, SINRs in (7), (8), (10) and (12) at high Δ can be approximated as follows:

$$\psi_{\mathrm{SR}}^{\mathrm{X}} \overset{\Delta\to+\infty}{\approx} \frac{\mu_{\mathrm{X}} g_{\mathrm{SR}}}{\sum_{k=1}^{K} 0.25 g_{\mathrm{I}_k\mathrm{R}} + 1}, \psi_{\mathrm{RU}_m}^{\mathrm{X}} \overset{\Delta\to+\infty}{\approx} \frac{(1-\mu_{\mathrm{X}}) g_{\mathrm{RU}_m}}{\sum_{k=1}^{K} 0.25 g_{\mathrm{I}_k\mathrm{U}_m} + 1}, \quad (39)$$

where $\mathrm{X} \in \{\mathrm{Con}V, ProP\}$.

Next, we can observe from Figure 2 that the OP performance of ConV and ProP is better as increasing $N_{\max}$. It is due to the fact that with higher $N_{\max}$, the ground users have more opportunity to sufficiently collect encoded packets for the data recovery. It is worth noting from Figure 2 that when $N_{\max} = 6$, the performance of ConV and ProP is the same, because in this case $N_{\max} = G_{\min}$. On the other hand, when $N_{\max} = 7$ and 8, the ProP scheme outperforms the ConV scheme. Furthermore, for the FHS condition with $N_{\max} = 8$, the OP variation of both schemes follows the same trend as in the AS case; however, due to the harsher propagation environment, the OP values under FHS are noticeably higher. It is also seen that as increasing $N_{\max}$, the performance gap between ConV and ProP also increases. It is noted that the effect of the Fountain coding overhead ε can be equivalently interpreted through $G_{\min} = (1+\varepsilon)P$. For a fixed $G_{\min}$, increasing $N_{\max}$ relaxes the decoding constraint, leading to a reduction in OP [28], [34]. This reveals a trade-off between coding overhead and transmission constraint in Fountain-coded systems. Finally, Figure 2 presents that the 'Sim' results validate the 'Theory' ones.

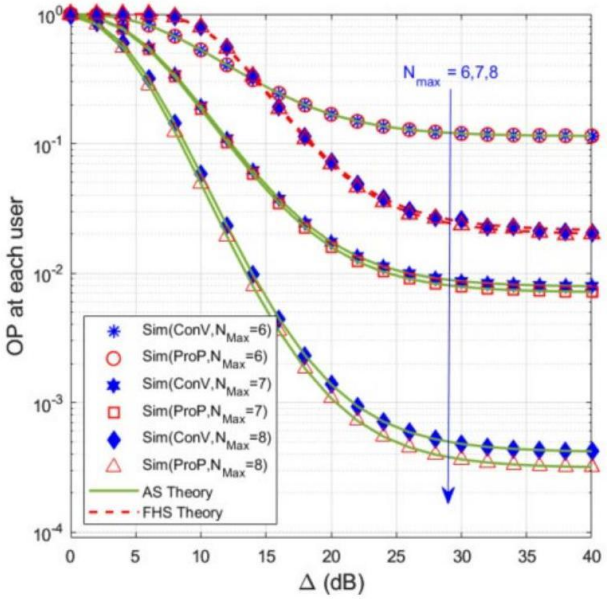


Figure 2. OP at each user as a function of Δ(dB) when $\mu_{\mathrm{ConV}} = \mu_{\mathrm{ProP}} = 0.5$ and $\tau_{\mathrm{ConV}} = \tau_{\mathrm{ProP}} = 0.5$.

Figure 3 presents OP at each user in ConV and ProP as a function of τ_{ConV} and τ_{ProP} when $\Delta = 20$ dB and $N_{\max} = 8$. As we can see from Figure 3, $\tau_{\mathrm{X}}(\mathrm{X} \in \{\mathrm{ConV}, \mathrm{ProP}\})$ significantly impacts the OP performance. Moreover, for each value of μ_{X}, there exists an optimal value of τ_{X} at which OP at each user in the X scheme is lowest. For example, with $\mu_{\mathrm{ConV}} = \mu_{\mathrm{ProP}} = 0.9$, OP at each user in ConV and ProP is minimized at $\tau_{\mathrm{ConV}} = 0.3$ and $\tau_{\mathrm{ProP}} = 0.35$, respectively. Figure 3 also illustrates that the OP performance of both ConV and ProP is almost best as $\mu_{\mathrm{ConV}} = \mu_{\mathrm{ProP}} = 0.5$. This is because when μ_{ConV} and μ_{ProP} are very low ($\mu_{\mathrm{ConV}} = \mu_{\mathrm{ProP}} = 0.1$), the transmit power of the satellite (S) is also very low. Conversely, when μ_{ConV} and μ_{ProP} are very high ($\mu_{\mathrm{ConV}} = \mu_{\mathrm{ProP}} = 0.9$), the transmit power of the relay station (R) becomes very low. These conditions result in a high OP at each user. Finally, Figure 3 again shows that ProP outperforms ConV, and the 'Sim' and 'Theory' results match very well.

In Figure 4, we present SOP of the ConV and ProP schemes as a function of Δ(dB) when $\mu_{\mathrm{ConV}} = \mu_{\mathrm{ProP}} = 0.5, \tau_{\mathrm{ConV}} = \tau_{\mathrm{ProP}} = 0.5$, and $N_{\max} = 8$. Similar to the OP at each user, the SOP values decrease as Δ increases, and converge to the saturation values at high Δ regimes. It is also seen that the SOP of ProP is lower that that of ConV. Moreover, the SOP performance of ConV and ProP is worse with higher number of ground users (M). Finally, it is worth noting that the 'Sim' results validate the 'Theory' ones.

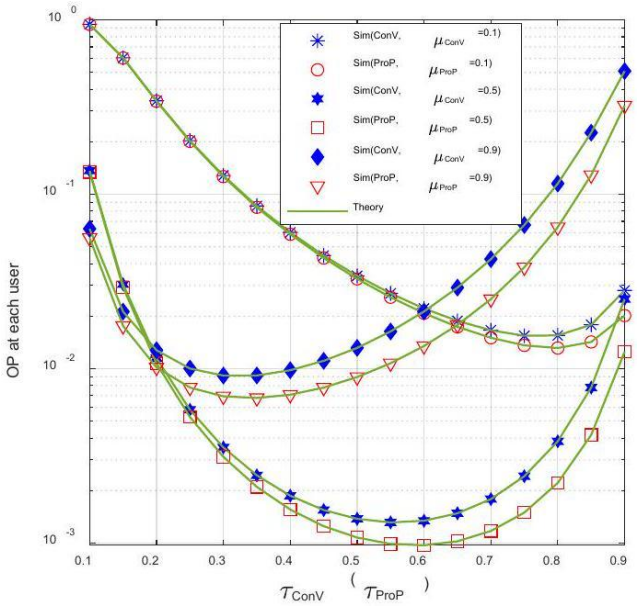


Figure 3. OP at each user as a function of τ_{ConV} (τ_{ProP}) when $\Delta = 20$ dB and $N_{\max} = 8$.

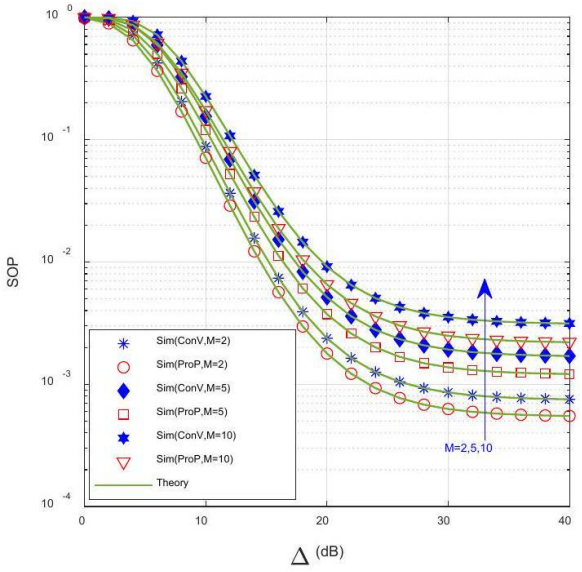


Figure 4. SOP as a function of Δ(dB) when $\mu_{\text{ConV}} = \mu_{\text{ProP}} = 0.5, \tau_{\text{ConV}} = \tau_{\text{ProP}} = 0.5$, and $N_{\max} = 8$.

Figure 5 presents SOP of the ConV and ProP schemes as a function of τ_{ConV} (τ_{ProP}) when $\Delta = 20(\text{ dB}), \mu_{\text{ConV}} = \mu_{\text{ProP}} = 0.65,$ and $M = 5$. We can see that the SOP performance of both ConV and ProP is significantly better when $N_{\max}$ increases from 7 to 8 . It is also observed that there exist optimal values of τ_{ConV} (τ_{ProP}) so that SOP of ConV (ProP) is lowest. For example, with $N_{\max} = 7$, SOP of ConV and ProP is lowest at $\tau_{\text{ConV}} = 0.4$ and $\tau_{\text{ProP}} = 0.45$, respectively. In addition, with $N_{\max} = 8$, SOP of ConV and ProP is lowest at $\tau_{\text{ConV}} = 0.45$ and $\tau_{\text{ProP}} = 0.5$, respectively. Similarly, the impact of ε on SOP is reflected through $G_{\min}$. A larger $G_{\min}$ requires more successfully received packets, whereas a larger $N_{\max}$ increases the decoding opportunity and hence reduces SOP.

Figure 6 investigates the impact of μ_{ConV} (μ_{ProP}) on the SOP performance of ConV and ProP with $\Delta = 15(\text{ dB}), M = 8,$ and $N_{\max} = 9$. As observed, there exist optimal values of μ_{ConV} and μ_{ProP}, so that SOP of ConV and ProP is lowest. For ConV, we can observe that SOP is lowest when $\tau_{\text{ConV}} = 0.55$ and $\mu_{\text{ConV}} = 0.5$, while ProP obtains the best performance when $\tau_{\text{ProP}} = 0.55$ and $\mu_{\text{ProP}} = 0.55$.

From Figures 5 and 6, it is worth noting that the joint time and power allocation problem (see (42)) must be solved to determine the optimal values of the ($\tau_{\text{x}}, \mu_{\text{x}}$) pairs, where $\text{X} \in \{\text{ConV}, \text{ProP}\}$. To achieve this, the Golden-section search algorithm presented in [42] can be employed.

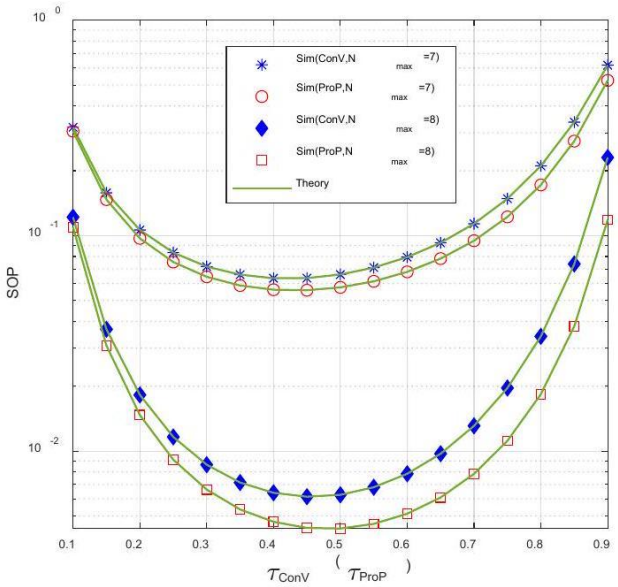


Figure 5. SOP as a function of τ_{ConV} (τ_{ProP}) when $\Delta = 20(\text{ dB}), \mu_{\mathrm{ConV}} = \mu_{\mathrm{ProP}} = 0.65$, and $M = 5$.

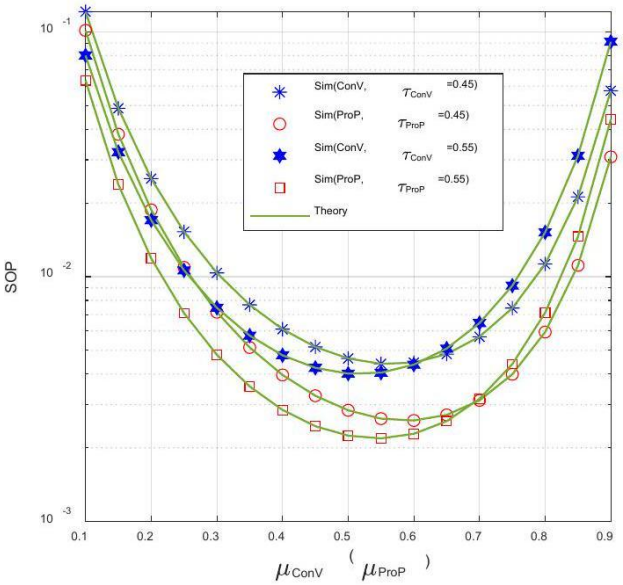


Figure 6. SOP as a function of μ_{ConV} (μ_{ProP}) when $\Delta = 15(\text{ dB}), M = 8$, and $N_{\max} = 9$.

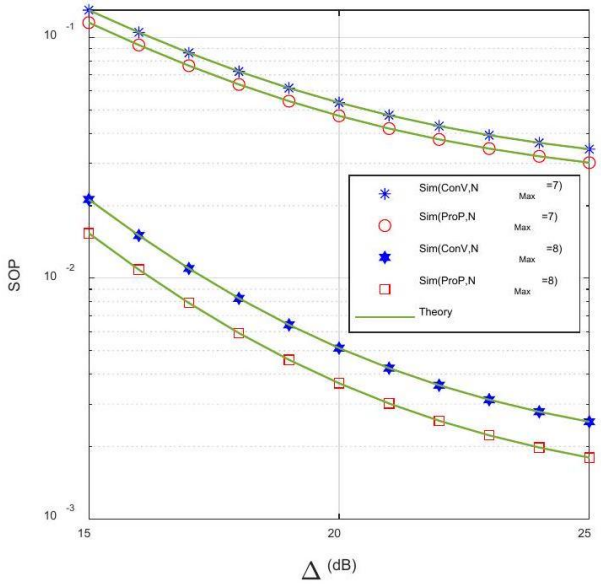


Figure 7. SOP as a function of $\Delta(\text{dB})$ with optimal values of ($\mu_{\mathrm{X}}, \tau_{\mathrm{X}}$) when $M = 5$.

Table 2. Optimal values of (μ_x, τ_x) in Figure 7 when $N_{\max} = 7$.

Δ	15 dB	16 dB	17 dB	18 dB	19 dB	20 dB	21 dB	22 dB	23 dB	24 dB	25 dB
μ^*_{ConV}	0.485	0.482	0.479	0.476	0.474	0.471	0.469	0.467	0.465	0.464	0.462
τ^*_{ConV}	0.501	0.499	0.496	0.494	0.491	0.489	0.487	0.485	0.483	0.482	0.481
μ^*_{ProP}	0.488	0.485	0.482	0.479	0.477	0.474	0.472	0.470	0.468	0.466	0.464
τ^*_{ProP}	0.512	0.510	0.507	0.504	0.502	0.499	0.497	0.495	0.494	0.492	0.491

Figure 7 presents SOP as a function of Δ in dB with various values of $N_{\max}$ and with $M = 5$. In this figure, the (μ_X, τ_X) pair is optimized according to equation (38). Indeed, Tables 2 and 3 present the optimal values of (μ_X, τ_X) in the case where $N_{\max} = 7$ and $N_{\max} = 8$, respectively. For example, in Table 2, with $N_{\max} = 7$ and $\Delta = 20$ dB, SOP of the ConV and ProP schemes is lowest at $(\mu_{\text{ConV}}, \tau_{\text{ConV}}) = (0.471, 0.489)$ and $(\mu_{\text{ProP}}, \tau_{\text{ProP}}) = (0.477, 0.502)$. For another example, in Table 3, with $N_{\max} = 8$ and $\Delta = 20$ dB, the optimal values of (μ_X, τ_X) are given as follows: $(\mu_{\text{ConV}}, \tau_{\text{ConV}}) = (0.503, 0.518)$ and $(\mu_{\text{ProP}}, \tau_{\text{ProP}}) = (0.509, 0.540)$. As seen from Figure 7, SOP of both ConV and ProP significantly decreases as increasing $N_{\max}$ and Δ. In addition, the performance gap between ConV and ProP also increases as $N_{\max}$ increases from 7 to 8.

Table 3. Optimal values of (μ_X, τ_X) in Figure 7 when $N_{\max} = 8$.

Δ	15 dB	16 dB	17 dB	17 dB	19 dB	20 dB	21 dB	22 dB	23 dB	24 dB	25 dB
μ^*_{ConV}	0.515	0.513	0.510	0.508	0.505	0.503	0.501	0.498	0.497	0.495	0.494
τ^*_{ConV}	0.529	0.527	0.525	0.522	0.520	0.518	0.516	0.514	0.512	0.511	0.509
μ^*_{ProP}	0.522	0.520	0.517	0.514	0.512	0.509	0.507	0.504	0.502	0.501	0.499
τ^*_{ProP}	0.551	0.549	0.547	0.544	0.542	0.540	0.537	0.536	0.534	0.533	0.531

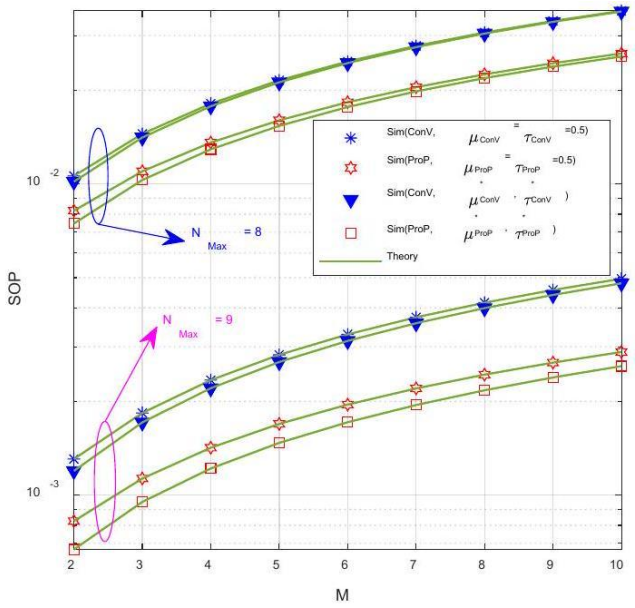


Figure 8. SOP as a function of M when $\Delta = 15$(dB).

Figure 8 presents the SOP performance of the considered schemes as a function of the number of ground users (M) with $\Delta = 15$(dB) and with two cases: i) the optimal values of (μ_X, τ_X); and $(\mu_X, \tau_X) = (0.5, 0.5)$. To simplify the presentation, the optimal values of the (μ_X, τ_X) pairs in Figure 8 will not be presented. As seen from Figure 8, SOP of both ConV and ProP increases as M increases. It is due to the fact that with higher number of ground users, the probability that at least one user experiences an outage, resulting in a higher SOP. Again, we can see that the SOP performance can be significantly enhanced by increasing $N_{\max} = 8$. In addition, ProP outperforms ConV, and the performance gap between ConV and ProP increases as increasing $N_{\max}$.

5. CONCLUSIONS

In this paper, we derived exact closed-form expressions of OP and SOP for both ConV and ProP schemes. These expressions were validated through Monte-Carlo simulations. Based on the derived SOP, we conducted the joint time and power allocation. The results demonstrate that the proposed scheme (ProP) outperforms the conventional scheme (ConV), in terms of both OP and SOP. Moreover, the performance gap in SOP between the two schemes increases as the number of transmission times ($N_{\max}$) increases. The findings also indicate that the SOP performance of ConV and ProP can be further improved by optimizing the time and power allocation parameters and by increasing $N_{\max}$. However, it is important to note that increasing $N_{\max}$ also leads to higher delay time and power consumption.

ACKNOWLEDGEMENTS

This work is supported by Posts and Telecommunications Institute of Technology in 2026.

REFERENCES

[1] S. Chen, S. Sun and S. Kang, "System Integration of Terrestrial Mobile Communication and Satellite Communication: Trends, Challenges and Key Technologies in B5G and 6G," China Communications, vol. 17, no. 12, pp. 156-171, 2020.

[2] E. Cianca et al., "Integrated Satellite-HAP systems", IEEE Communications Magazine, vol. 43, no. 12, pp. supl.33-supl.39, 2005.

[3] K. An and T. Liang, "Hybrid Satellite-terrestrial Relay Networks with Adaptive Transmission", IEEE Transactions on Vehicular Technology, vol. 68, no. 12, pp. 12448-12452, 2019.

[4] X. Li, W. Feng, J. Wang, Y. Chen, N. Ge and C. -X. Wang, "Enabling 5G on the Ocean: A Hybrid Satellite-UAV-terrestrial Network Solution," IEEE Wireless Comm., vol. 27, no. 6, pp. 116121, 2020.

[5] K. Mashiko et al., "Combined Control of Coverage Area and HAPS Deployment in Hybrid FSO/RF SAGIN," IEEE Transactions on Vehicular Technology, vol. 74, no. 7, pp. 10819-10828, 2025.

[6] S. Yuan, Y. Sun and M. Peng, "Joint Network Function Placement and Routing Optimization in Dynamic Software-defined Satellite-terrestrial Integrated Networks," IEEE Transactions on Wireless Communications, vol. 23, no. 5, pp. 5172-5186, DOI: 10.1109/TWC.2023.3324729, May 2024.

[7] S. Yuan, Y. Sun, M. Peng and R. Yuan, "Joint Beam Direction Control and Radio Resource Allocation in Dynamic Multi-Beam LEO Satellite Networks," IEEE Transactions on Vehicular Technology, vol. 73, no. 6, pp. 8222-8237, DOI: 10.1109/TVT.2024.3353339, June 2024.

[8] S. Yuan, Y. Sun and M. Peng, "Cache-aware Cooperative Multicast Beamforming in Dynamic Satellite-terrestrial Networks," IEEE Transactions on Vehicular Technology, vol. 74, no. 1, pp. 1433-1445, 2025.

[9] S. Yuan, M. Peng and Y. Sun, "Satellite-terrestrial Integrated Fog Networks: Architecture, Technologies, and Challenges," IEEE Wireless Communications, vol. 32, no. 4, pp. 208-215, August 2025.

[10] C. Ding, J. -B. Wang, H. Zhang, M. Lin and G. Y. Li, "Joint MIMO Precoding and Computation Resource Allocation for Dual-function Radar and Communication Systems with Mobile Edge Computing," IEEE Journal on Selected Areas in Communications, vol. 40, no. 7, pp. 2085-2102, July 2022.

[11] B. Zhao, M. Lin, B. Ma, J. Ouyang, N. Al-Dhahir and M. -S. Alouini, "LDM-based Communication and Computation Co-design in Integrated Satellite and Aerial Networks," IEEE Transactions on Communications, vol. 73, no. 11, pp. 10230-10245, DOI: 10.1109/TCOMM.2025.3568218, Nov. 2025.

[12] Q. Huang et al., "Secrecy Performance of Hybrid Satellite-Terrestrial Relay Networks in the Presence of Multiple Eavesdroppers," IET Communications, vol. 12, no. 1, pp. 26-34, 2018.

[13] W. Cao, Y. Zou, Z. Yang and J. Zhu, "Relay Selection for Improving Physical-layer Security in Hybrid Satellite-terrestrial Relay Networks," IEEE Access, vol. 6, pp. 65275-65285, 2018.

[14] W. Cao et al., "Security-reliability Trade-off Analysis of Hybrid Satellite-terrestrial Uplink Communications with Relay Selection," IEEE Systems Journal, vol. 18, no. 1, pp. 485-496, 2024.

[15] K. Guo, K. An, B. Zhang, Y. Huang and G. Zheng, "Outage Analysis of Cognitive Hybrid Satellite-terrestrial Networks with Hardware Impairments and Multi-primary Users," IEEE Wireless Communications Letters, vol. 7, no. 5, pp. 816-819, 2018.

[16] V. Singh, S. Solanki and P. K. Upadhyay, "Cognitive Relaying Cooperation in Satellite-terrestrial Systems with Multiuser Diversity," IEEE Access, vol. 6, pp. 65539-65547, 2018.

[17] Y. Guo, M. Lin, Y. Liu, H. Kong, J. -B. Wang and J. Wang, "AoI-aware Uplink CR-NOMA Schemes in Satellite Internet of Things Networks," IEEE Transactions on Aerospace and Electronic Systems, vol. 61, no. 1, pp. 1224-1230, DOI: 10.1109/TAES.2024.3451455, Feb. 2025.

[18] V. Singh, P. K. Upadhyay, D. B. da Costa and U. S. Dias, "Hybrid Satellite-terrestrial Spectrum Sharing Systems with RF Energy Harvesting," Proc. of 2018 IEEE 29th Annual Int. Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), pp. 306-311, Bologna, Italy, 2018.

[19] Z. Li, G. Wang and M. Yang, "Performance Analysis of SWIPT Aided Satellite-terrestrial Cooperative Network," Proc. of the 2nd Asia-Pacific Conf. on Communications Technology and Computer Science (ACCTCS), pp. 252-256, Shenyang, China, 2022.

[20] V. Singh and P. K. Upadhyay, "Exploiting FD/HD Cooperative-NOMA in Underlay Cognitive Hybrid Satellite-terrestrial Networks," IEEE Transactions on Cognitive Communications and Networking, vol. 8, no. 1, pp. 246-262, 2022.

[21] T. N. Nguyen et al., "Outage Performance of Satellite Terrestrial Full-duplex Relaying Networks with Co-channel Interference," IEEE Wireless Communications Letters, vol. 11, no. 7, pp. 1478-1482, 2022.

[22] Z. Lin et al., "Refracting RIS-aided Hybrid Satellite-terrestrial Relay Networks: Joint Beamforming Design and Optimization," IEEE Transactions on Aerospace and Electronic Systems, vol. 58, no. 4, pp. 3717-3724, 2022.

[23] X. Yan, H. Xiao, C.-X. Wang and K. An, "Outage Performance of NOMA-based Hybrid Satellite-terrestrial Relay Networks", IEEE Wireless Communications Letters, vol. 7, no. 4, pp. 538-541, 2018.

[24] L. Han, W.-P. Zhu and M. Lin, "Outage of NOMA-based Hybrid Satellite-terrestrial Multi-antenna DF Relay Networks," IEEE Wireless Communications Letters, vol. 10, no. 5, pp. 1083-1087, 2021.

[25] V. Singh, V. Bankey and P. K. Upadhyay, "Underlay Cognitive Hybrid Satellite-terrestrial Networks with Cooperative-NOMA," Proc. of 2020 IEEE Wireless Communications and Networking Conf. (WCNC), pp. 1-6, Seoul, Korea, 2020.

[26] H. -N. Nguyen et al., "Reliable and Secure Transmission in Multiple Antennas Hybrid Satellite-terrestrial Cognitive Networks Relying on NOMA," IEEE Access, vol. 8, pp. 215044-215056, 2020.

[27] L. Han, W.-P. Zhu and M. Lin, "Outage Analysis of Multi-relay NOMA-based Hybrid Satellite-terrestrial Relay Networks," IEEE Transactions on Vehicular Technology, vol. 71, no. 6, pp. 64696487, 2022.

[28] D. J. C. MacKay, "Fountain Codes," IEE Proceedings-Communications, vol. 152, no. 6, pp. 10621068, 2005.

[29] T. L. Thanh et al., "Broadcasting in Cognitive Radio Networks: A Fountain Codes Approach," IEEE Transactions on Vehicular Technology, vol. 71, no. 10, pp. 11289-11294, 2022.

[30] N. V. Toan et al., "Outage Performance of Hybrid Satellite-terrestrial Relaying Networks with Rateless Codes in Co-channel Interference Environment," Proc. of 2023 Int. Conf. on System Science and Engineering (ICSSE), pp. 468-473, Ho Chi Minh, Vietnam, 2023.

[31] 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, 2023.

[32] P. M. Quang et al., "Performance Enhancement for Rateless Codes-aided Hybrid Satellite-terrestrial Multi-user Networks Using NOMA and IRS with Presence of Multiple Eavesdroppers," Proc. of the 9th Int. Conf. on Consumer Electronics Asia (ICCE-Asia), pp. 1-4, Danang, Vietnam, 2024.

[33] N. V. Toan et al., "Performance Evaluation of Hybrid Satellite-terrestrial Relaying Broadcast Networks Using Fountain Codes and NOMA," Proc. of 2024 IEEE Int. Conf. on Consumer Electronics-Asia (ICCE-Asia), pp. 1-4, Danang, Vietnam, 2024.

[34] N. V. Toan, T. T. Duy, P. N. Son, P. V. Tuan and L. T. Tu, "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 Transactions on Industrial Networks and Intelligent Systems, vol. 12, no. 03, pp. 1-11, 2025.

[35] L. Han, W.-P. Zhu, and M. Lin, "Uplink outage performance of NOMA-based hybrid satellite-terrestrial relay networks over generalized inhomogeneous fading channels," IEEE Transactions on Communications, vol. 70, no. 4, pp. 2417-2434, 2022.

[36] N. Q. Sang et al., "Securing Wireless Communications with Energy Harvesting and Multi-antenna Diversity", Jordanian Journal of Computers and Information Technology (JJCIT), vol. 11, no. 02, pp. 197-210, June 2025.

[37] D.-H. Ha, T. T. Duy, P. N. Son, T. Le-Tien and M. Voznak, "Security-reliability Trade-off Analysis for Rateless Codes-based Relaying Protocols Using NOMA, Cooperative Jamming and Partial Relay Selection," IEEE Access, vol. 9, pp. 131087-131108, 2021.

[38] T. N. Nguyen et al., "Outage Performance of Satellite Terrestrial Full-duplex Relaying Networks with Co-channel Interference," IEEE Wireless Communications Letters, vol. 11, no. 7, pp. 1478-1482, 2022.

[39] N. Q. Sang et al., " Power Beacon-assisted Energy Harvesting in D2D Network under Co-channel Interferences: Symbol Error Rate Analysis ", Jordanian Journal of Computers and Information Technology (JJCIT), vol. 11, no. 04, pp. 517-532, December 2025.

[40] A.-T. Le et al., "Physical Layer Security Analysis for RIS-aided NOMA Systems with Non-colluding Eavesdroppers," Computer Communications, vol. 219, pp. 194-203, 2024.

[41] B. Li, Y. Zou, T. Wu, Z. Zhang, M. Chen and Y. Jiang, "Security and Reliability Tradeoff of NOMA Based Hybrid Satellite-terrestrial Network with a Friendly Jammer," IEEE Transactions on Vehicular Technology, vol. 74, no. 2, pp. 3439-3444, Feb. 2025.

[42] E. K. P. Chong and S. H. Zak, An Introduction to Optimization, DOI: 10.1002/9781118033340, ISBN: 9780471758006, United States: Wiley, 2008.

ملخص البحث:

نـدرس فـي هـذه الورقـة أداء الانقطـاع لأنظمـة البـث المتعـدّد الهجينـة عبـر الأقمـار الصّـناعية والشّـبكات الأرضـية باسـتخدام رمـوز (Fountain). فـي المخطّطـات المدروسـة، يحـاول القمـر الصّـناعي إرسـال بياناتـه إلـى مجموعـة مـن المسـتخدمين الأرضـيّين بمسـاعدة محطّـة تقويـة أرضـية. فـي المخطـط التّقليـدي (ConV)، تقـوم محطّـة التّقويـة بإعـادة توجيـه كـل حزمـة بيانـات (Fountain) إلـى المسـتخدمين الأرضيّين باستخدام تقنية فكّ التّشفير وإعادة التّوجيه.

أمّـا فـي المخطّـط المقتـرح (PropP)، فتقـوم محطّـة التّقويـة بتخـزين حـزم بيانـات (Fountain) المسـتلمة مـن القمـر الصّـناعي، وتقـوم مقـام القمـر الصّـناعي فـي إرسـال حـزم بيانـات (Fountain) جديـدة إلـى المسـتخدمين الأرضـيّين بمجـرّد جمّـع عـددٍ كـافٍ مـن حـزم بيانـات (Fountain) لاسـتعادة البيانـات. ونسـتنتج صِـيغاً مغلقـة دقيقـة لاحتماليـة انقطـاع الطّاقـة عنـد كـل مسـتخدم، واحتماليـة انقطـاع الخدمـة للمخطّطـين المـذكورين (التّقليدي والمقترح) مع مراعاة تأثير التّداخل بين القنوات المتجاورة.

وقـد تـمّ إجـراء محاكـاة حاسـوبية للتّحقُّـق مـن صـحّة الصِّـيغ المسـتخدمة. عـلاوةً علـى ذلـك، تمّـت صـياغة مسـألة مشـتركة وحلّهـا لتخصـيص الوقـت والطّاقـة معـاً لتحسـين أداء احتمالية انقطاع الخدمة للنّظام في المخطّطين المدروسين.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

FIXED-SET LEARNING FOR CLUSTER-HEAD SELECTION IN MULTI-HOP WIRELESS SENSOR NETWORKS

Raouf Ouanis Lakehal Ayat[1] and Salim Bouamama[2]

(Received: 16-Feb.-2026, Revised: 7-Apr.-2026, Accepted: 30-Apr.-2026)

ABSTRACT

Wireless Sensor Networks (WSNs) have remained an active research field in both military and civilian domains, driven by the expanding diversity of their applications. In recent years, there has been a progressive shift toward integrating Artificial Intelligence to address the persistent challenge of energy optimization in WSNs. We introduce a novel adaptation of a Fixed Set Search (FSS) mechanism to WSNs. FSS adds a learning phase to the well-known GRASP metaheuristic. FSS-WSN approach guides the Base Station (BS) in a centralized multi-hop environment to select the optimal cluster-heads, thereby maximizing the global utility of the network. We evaluated our approach under documented fairness conditions, against a wide range of established baselines including classical clustering protocols (LEACH, HEED, SEP), widely used swarm optimizers (PSO, GWO, ABC), the recent multi-hop hybrid EEM-LEACH-ABC, and recent SO-GJO-family variants (SO, GJO, EMO–GJO, and ESO–GJO). The results demonstrate a statistically significant improvement (paired Wilcoxon test with Holm correction) over the best baseline regarding two key metrics–the number of delivered reports and the CPU time required for decision-making. These results suggest that our approach is a strong, practical option for many WSN use cases.

KEYWORDS

Wireless sensor networks, Cluster-head selection, Multi-hop, Fixed Set Search, Metaheuristics, Energy efficiency.

1. INTRODUCTION

Energy remains a major limitation in most Wireless Sensor Networks (WSNs), particularly in remote or safety-critical deployments where replacing batteries is impractical. On many platforms, radio communication accounts for the largest share of energy consumption. Consequently, both network lifetime and the volume of data successfully delivered to the Base Station (BS) are closely tied to the way in which measurements are aggregated and forwarded [1] [2] [3].

WSN-based IoT architectures have been deployed in diverse safety-critical domains, including gas-leakage detection [4], smart-grid monitoring [5], and industrial-process control [1]. In all such settings, energy-efficient clustering directly impacts system reliability and data availability.

To address this issue, clustering has become a common approach. Sensor nodes transmit short-range reports to a cluster head (CH), which aggregates the data and forwards it to the BS, thereby reducing the number of long-distance transmissions. To avoid premature energy depletion of a small subset of nodes, most clustering protocols periodically rotate the CH role, as in LEACH [6] and HEED [7]. In networks with heterogeneous initial energy levels, CH selection is often weighted in favor of higher-energy nodes, as in SEP [8].

Beyond extending network lifetime, many WSN applications place strong emphasis on the timely and reliable delivery of decision-relevant information. In scenarios, such as industrial monitoring, emergency response, or perimeter surveillance, the objective is not only to keep the network operational, but also to maximize the amount of useful data delivered under strict time and energy constraints.

Things get more complicated when multi-hop forwarding enters the picture, because coverage and routing can no longer be treated separately. The set of CHs chosen in a given round does not just decide which sensor nodes are covered—it also shapes the relay chain that funnels data back to the BS. In practice, CHs that sit near the sink often end up relaying traffic for distant clusters on top of their own, which drains them faster and creates localized energy bottlenecks [9] [2]. Most existing formulations gloss over this coupling, either by using rough distance-based surrogates or by folding feasibility into penalty terms that the optimizer can partially ignore.

1. R. O. Lakehal Ayat is with Department of Computer Science, University of M'sila, Algeria. Email: raouf.lakehalayat@univ-msila.dz
2. S. Bouamama is with Department of Computer Science, University of Setif 1–Ferhat Abbas, Sétif 19000, Algeria. Email: salim.bouamama@univ-setif.dz

1.1 Related Work

Over the past two decades, a large number of protocols and algorithms have been proposed for cluster-head selection and clustered routing in WSNs. Early work focused on lightweight distributed heuristics, while more recent studies have increasingly turned to optimization-based formulations. We review the most relevant contributions below, grouped by theme.

- **Energy-sensitive clustering protocols:** A handful of distributed protocols still set the reference point for CH election. LEACH [6] introduced probabilistic self-election with round-by-round role rotation and local data aggregation—a design that remains surprisingly competitive given its simplicity. HEED [7] took a slightly different route, iterating over both residual energy and intra-cluster communication cost; the result, in our experience, is noticeably more balanced cluster sizes. For networks where batteries are not identical, SEP [8] provides a natural fix by scaling each node's election weight with its initial energy. A more recent entry is EEM-LEACH-ABC [10], in which Zhang et al. use an Artificial Bee Colony optimizer to automatically set the CH ratio and an energy-blending parameter within a LEACH-like framework. While their numbers look favorable, the study covers only one network layout under fixed settings and does not include replicated experiments, which makes it difficult to draw firm conclusions. It is worth stressing that every protocol in this group is distributed by design; we therefore treat them as reference baselines and not as direct alternatives to the centralized, multi-hop optimization that we develop.

- **Metaheuristic-based CH selection:** CH selection is often framed as an optimization problem, typically involving a weighted combination of energy, distance, and load balance [11], on the face of it, a sensible idea, and a long list of continuous-space metaheuristics have been applied to it: PSO [12], GWO [13], ABC [14], SO [15], GJO [16], among others. Some of these are growing more sophisticated. Gupta et al., for instance, fold relay-load balancing into a GSO-based clustering framework [17] and tackle path planning for heterogeneous WSNs separately [18]. All of this sounds promising, yet the numbers are hard to trust across studies: change in the continuous-to-discrete decoding rule or tweak coverage feasibility checking, and the rankings can shift substantially [19].

- **Augmented metaheuristic frameworks:** A few recent studies have tried to patch the encoding and feasibility problems by bolting domain-specific modules onto existing metaheuristics. Mazumder et al. recast CH selection as a multi-objective energy-aware task and attacked it with a Golden Jackal Optimizer variant which they call EMO-GJO [20]—an interesting formulation, though the evaluation stays within a single simulator setup. Wang et al. [21] went further: their ESO–GJO couples the Snake Optimizer with GJO, so that CH placement and multi-hop routing are decided in one pass. Both papers show energy and lifetime gains, but since each team ran its own simulator with its own parameter conventions, putting their numbers side by side with anything else is not straightforward.

Structural limitations of continuous-space approaches: Strip away the algorithmic labels and every metaheuristic reviewed above—plain or augmented—does the same thing: search a continuous $[0,1]^N$ space, then snap the result to a binary CH vector with a threshold or top-k rule. That two-step detour has real consequences. Huge swaths of the continuous space collapse onto the same CH set, so the optimizer wastes iterations exploring what is effectively one solution. Nudge a single coordinate by a small amount and a node can flip in or out of the CH set, leaving the fitness surface jagged in ways the search operators were never designed for. Worse, the memory each algorithm carries—PSO's velocities and personal bests, the alpha–beta–delta hierarchy in GWO, ABC's food-source vectors—lives entirely in continuous coordinates. None of it records which nodes tend to show up in good CH configurations.

Table 1 summarizes these structural differences. Other memory-driven paradigms fare little better: ACO pheromone trails and Tabu Search lists [22] [23] track individual moves, not recurring sub-sets of promising nodes.

1.2 Gap and Contributions

In the round-based centralized setting that most metaheuristic studies adopt, the BS picks a CH set at the start of each round and broadcasts the assignment to all nodes. As discussed above and summarized in Table 1, every existing optimizer works in a continuous space and must decode its solutions into binary CH vectors—a process that severs the link between the algorithm's internal memory and the discrete structure of the problem that it is actually solving. What is missing, in particular, is a mechanism

that identifies which specific nodes keep showing up in good CH configurations during a given round and feeds that information back into the search.

Table 1. How FSS-WSN differs structurally from continuous-space CH selection methods.

Feature	PSO, GWO, ABC, SO, GJO variants	FSS-WSN (this work)
Search space	Real-valued $[0,1]^N$	Combinatorial (CH subsets built directly)
Decoding	A threshold or top-*k* step is always needed to obtain a binary CH vector	Skipped entirely: solutions are already discrete
What the memory captures	Continuous coordinates: velocity & personal bests (PSO), pack hierarchy (GWO), food-source positions (ABC) none tied to specific nodes	Which individual nodes recur in high-quality CH configurations
Exploitation style	Population converges toward a global-best real-valued vector	Later GRASP iterations are biased by the fixed set of promising CHs
Feasibility	Usually enforced *via* penalty terms; *post-hoc* repair added in some variants	Hard two-stage repair (coverage then connectivity), with a regularization term penalizing solutions that rely heavily on it
Per-round CPU*	1.1–2.1 s (varies across algorithms)	0.07–0.09 s

*AMD Ryzen 5 7600X (6C/12T), Python 3.12, N=100; see Subsection 4.2.

We fill this gap by adapting Fixed Set Search (FSS) [24] [25], a hybrid metaheuristic designed for combinatorial problems, to centralized CH selection in multi-hop WSNs where routing is part of the decision. The main contributions are as follows:

1) **Native combinatorial search:** FSS-WSN works directly on CH sub-sets, without any continuous-to-binary conversion. Each candidate solution is built by a coverage-driven GRASP procedure that enforces the cluster radius during construction, avoiding the selection rules commonly used in continuous-space methods.
2) **Two-phase in-round search with component-level memory:** Each round is split into two phases. Phase I runs a diversified GRASP to build a pool of high-quality CH sets; from this pool, per-node selection frequencies are computed and a fixed set of consistently promising CH candidates is extracted. Phase II then biases new constructions toward nodes in the fixed set, provided that they still have sufficient energy (an energy-guard check prevents the algorithm from over-exploiting depleted nodes). Because the fixed set is rebuilt at every round, it naturally tracks the shifting energy landscape.
3) **Deterministic repair with regularization:** Feasibility is enforced through a two-stage deterministic procedure: the first stage ensures that every sensor is covered within the cluster radius, and the second ensures that every CH can reach the sink within its transmission range. No randomness is involved, so results are fully reproducible. A regularization term in the objective penalizes solutions that lean heavily on repair, pushing the search toward configurations that are feasible by construction.
4) **Uniform experimental comparison:** All 11 baselines—PSO, GWO, ABC, SO, GJO, EMO-GJO, ESO-GJO, LEACH, HEED, SEP, and EEM-LEACH-ABC are reimplemented within a single Python simulator that enforces the same energy model, repair procedure, multi-hop Dijkstra routing, and metric definitions. Every experiment is replicated over 30 paired random seeds and assessed with Wilcoxon signed-rank tests under Holm correction to control family-wise error.
5) **Best throughput with markedly lower CPU cost:** FSS-WSN achieves the highest cumulative throughput in all four tested scenarios (homogeneous/heterogeneous energy with center/corner BS), with Holm-corrected $p<0.05$ over every baseline. Per-round computation is 10–25 times faster than the continuous-space alternatives, a margin relevant for deployment scenarios where CH assignments must be recomputed in near-real time.

1.3 Paper Organization

The remainder of this paper is organized as follows. Section 2 introduces the system model and the

routing-coupled problem formulation. Section 3 describes the proposed FSS-WSN method in detail. Sections 4 and 5 present the evaluation protocol and discuss the experimental results. Finally, Section 6 concludes the paper.

2. System Model and Problem Formulation

2.1 Two-graph Abstraction

Notation: In this subsection, V_t denotes the set of alive nodes and $d(u, v)$ the Euclidean distance between any two nodes u and v. Edge sets are denoted by E_t^{cov} and E_t^{tx}.

To model clustering and routing at round t, we use two distinct graphs: one for cluster membership and one for multi-hop forwarding. This separation makes the execution rules deterministic, i.e., under identical conditions, the same decisions are obtained. As in classical WSN models, coverage is governed by a fixed clustering radius R_c, while data forwarding is restricted to a cluster-head (CH) backbone. Energy dissipation follows the first-order radio model [6] [26]. At each round, CH selection, routing, and constraint satisfaction are evaluated under a fixed decision budget.

We use two thresholds:

- R_c (Coverage radius): governs cluster membership (clustering feasibility).
- r_{tx} (Transmission radius): governs one-hop connectivity for routing.

Figure 1 illustrates this distinction:

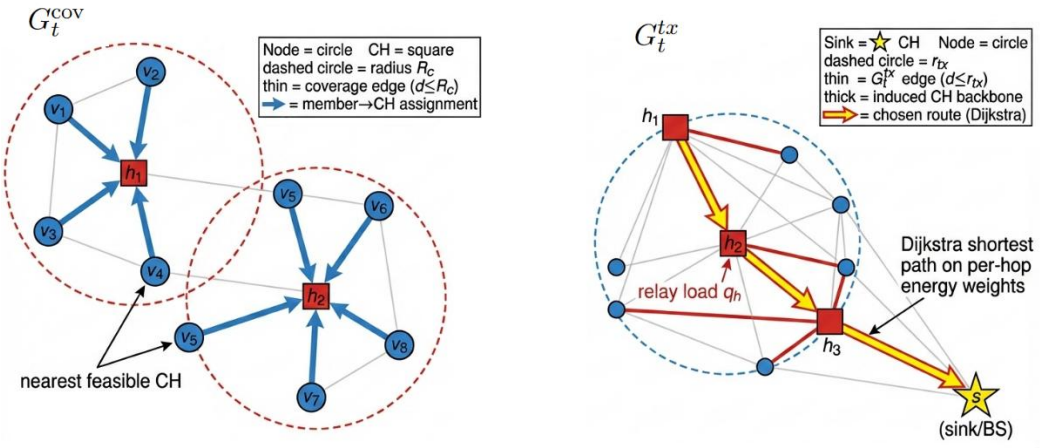


Figure 1. Graph abstractions used in this work.

In all experiments, we set $R_c = 25\,\text{m}$ and $r_{tx} = 50\,\text{m}$ (i.e., $r_{tx} = 2R_c$) unless stated otherwise. The clustering (coverage) graph is $G_t^{\text{cov}} = (V_t, E_t^{\text{cov}})$, where $(u, v) \in E_t^{\text{cov}}$ if and only if $d(u, v) \le R_c$. The communication graph is $G_t^{tx} = (V_t, E_t^{tx})$, where $(u, v) \in E_t^{tx}$ if and only if $d(u, v) \le r_{tx}$.

Keeping these constraints separate is important: full coverage does not guarantee balanced multi-hop forwarding. A design may still create relay bottlenecks (hotspot) and drain the nodes closest to the sink, this phenomenon is known as the 'energy hole' problem [27] especially when the sink placement is unfavorable (e.g., at the corner) [2] [9].

In round t, a candidate is a CH seed set denoted $\mathcal{H}_0 \subseteq V_t$. We enforce feasibility *via* the deterministic mapping to construct the repaired CH set $\mathcal{H}^+ = \text{Repair}_t(\mathcal{H}_0)$, and only $\mathcal{H}^+$ is deployed. The repair step guarantees a non-empty CH set, strict radius coverage.

$\text{Repair}_t(\cdot)$ proceeds deterministically in two stages:

First, in the coverage: if $\mathcal{H}_0 = \varnothing$, it selects one initial CH maximizing a fixed greedy score (residual-energy ratio, sink-centrality, local density; ties by smallest node index), then iteratively adds the node that covers the largest number of currently uncovered alive nodes under R_c (ties by smallest index) until (1) holds.

Second, in the connectivity (multi-hop): if no CH is within r_{tx} of the sink, it promotes the highest-energy sink-direct node as a CH, then connects remaining CHs by promoting intermediate nodes along a

deterministic BFS path on G_t^{tx} to a connected component containing a sink-direct CH (neighbors explored in increasing index order).

$$\forall v \in V_t,\ \exists h \in \mathcal{H}^+ \ s.t.\ d(v,h) \le R_c \tag{1}$$

and CH-to-sink reachability over the inter-CH backbone under r_{tx}. Each node then attaches to its nearest CH in $\mathcal{H}^+$ (using deterministic tie-breaking). This repair-first convention is standard in constrained optimization [19]; moreover, minimizing $|\mathcal{H}^+|$ subject to (1) reduces to the NP-complete minimum dominating-set problem [28], for which approximation algorithms with logarithmic ratios have been established [29].

2.2 Bounded Fitness Surrogate

A link (u, v) exists if and only if $d(u, v) \le r_{tx}$, i.e., u and v can communicate in one hop. We weight each edge by the energy spent to forward one aggregated packet of length L over that hop using the first-order radio model [6] [26], which is widely adopted in the WSN literature for energy accounting. This choice makes the routing cost physically meaningful: it captures the strong dependence of transmission energy on distance and thus reflects the actual energy drain induced by multi-hop forwarding. Running Dijkstra on the graph $G^{mh}(\mathcal{H}^+)$ (the multi-hop backbone derived from the global communication graph G_t^{tx}) then yields, for each CH $h \in \mathcal{H}^+$, the minimum-energy delivery cost κ_h to the sink s.

Score normalization: All surrogate components with the symbol tilde (e.g., $\tilde{C}_D$, $\tilde{C}_L$) are normalized to lie in [0,1] (lower is better). We use $\mathrm{clip}_{[0,1]}(x)$ to truncate any value x to the interval [0,1].

Bounded surrogate terms and component definitions We compute a bounded surrogate in the repaired clustering that combines: CH energy robustness C_E, member-to-CH distance $\tilde{C}_D$, cluster-size imbalance $\tilde{C}_L$, and a routing-aware term $\tilde{C}_S^{mh}$ derived from $\{\kappa_h\}$. Specifically:

- C_E is the average CH energy depletion ratio $1 - E_{\mathrm{res}}/E_{\mathrm{init}}$ over $h \in \mathcal{H}^+$;
- $\tilde{C}_D$ is the mean distance from each alive node to its selected (nearest) CH in $\mathcal{H}^+$, normalized by the area diagonal;
- $\tilde{C}_L$ is the normalized variance of the resulting cluster sizes $\{|C_h|\}_{h\in\mathcal{H}^+}$ (load imbalance);
- $\tilde{C}_S^{mh}$ is the mean minimum-energy multi-hop delivery cost from each CH to the sink (Dijkstra on $G^{mh}(\mathcal{H}^+)$), normalized by a fixed constant.

From the Dijkstra next-hop structure, we also compute each CH relay load q_h (own plus relayed packets) and penalize relay-load dispersion *via* $\tilde{C}_R$, defined as the normalized variance of CH relay loads induced by the next-hop structure.

Base objective and anti-hotspot component:

We first define:

$$F_0 = w_1\, C_E + w_2 \left(\frac{\tilde{C}_D + \tilde{C}_S^{mh}}{2}\right) + w_3\, \tilde{C}_L, \qquad w_1 + w_2 + w_3 = 1. \tag{2}$$

and then incorporate the anti-hotspot term as:

$$F_{\mathrm{base}} = (1 - w_R)\, F_0 + w_R\, \tilde{C}_R, \qquad w_R \in [0,1]. \tag{3}$$

Repair-dependence regularizer and final fitness: While repair mechanisms guarantee feasibility, naive repair can bias metaheuristics toward 'easy-to-repair' regions of the search space—a phenomenon discussed in constrained optimization literature [19] [30]. To address this bias, we penalize candidates that rely heavily on repair, we measure the fraction of CHs added by $\mathrm{Repair}_t(\cdot)$, where $|\cdot|$ denotes set cardinality and "\" denotes set difference:

$$P(\mathcal{H}_0) = \frac{|\mathcal{H}^+ \backslash \mathcal{H}_0|}{N_t} \in [0,1]. \tag{4}$$

We measure the per-round regularized objective:

$$\mathrm{Fitness}_t(\mathcal{H}_0) = \mathrm{clip}_{[0,1]}(F_{\mathrm{base}}) + \lambda\, P(\mathcal{H}_0), \qquad \lambda > 0. \tag{5}$$

The parameter $\lambda > 0$ weights the repair-dependence penalty and discourages seeds that rely on $\mathrm{Repair}_t(\cdot)$.The BS keeps the candidate with minimum fitness, applies $\mathrm{Repair}_t(\cdot)$, and deploys the repaired CH set.

3. PROPOSED FSS-WSN ALGORITHM

3.1 Algorithm Description

At the beginning of each round t, the Base Station (BS) computes a deployable clustered configuration using **FSS-WSN**. Let $V_t = \{i \in V\colon E_i^{\mathrm{res}}(t) > 0\}$ denote the alive nodes at round start; $E_i^{\mathrm{res}}(t)$ denotes residual energy (should not be confused with the edge sets E_t^{cov}, E_t^{tx}, or E^{mh}). The decision variable is a *seed* set of prospective cluster heads (CHs), $\mathcal{H} \subseteq V_t$; the configuration ultimately deployed in the network is obtained by selecting the seed that minimizes the regularized surrogate fitness (Eq. (5)).

All seeds are evaluated under a fixed deployment convention. Given $\mathcal{H}$, we first apply the deterministic feasibility mapping $\mathcal{H}^+ = \mathrm{Repair}_t(\mathcal{H})$ (Eq. (4)). This mapping enforces three hard requirements: $\mathcal{H}^+ \neq \emptyset$; strict R_c coverage for member attachment; and reachability of each CH to the sink on the inter-CH backbone under r_{tx}.

Once $\mathcal{H}^+$ is obtained, every alive node attaches to its nearest CH, and multi-hop forwarding costs are computed by running Dijkstra on the induced CH backbone with energy-based hop weights. The deployed round output is therefore $\mathcal{H}_t^{\star+} = \mathrm{Repair}_t(\mathcal{H}_t^{\star})$, where $\mathcal{H}_t^{\star}$ is the best seed under $\mathrm{Fitness}_t(\cdot)$.

FSS-WSN follows a two-phase routine. Phase I populates a small elite pool using GRASP-style randomized greedy constructions (optionally followed by bounded swaps). Phase II extracts from that pool a compact set of recurrent, energy-safe CH nodes and uses this set as a warm start for additional constructions [24] [25] [31] [32].

Algorithm 1 summarizes the round-level procedure. (*Unless stated otherwise, swap refinement is disabled in experiments* $L_{\max} = 0$).

The algorithm follows a two-pass pattern:

- **Phase I (diversified GRASP seeds):** Starting from $\emptyset$, we build seed CH sets using GRASP until all alive nodes are covered within R_c or the construction reaches K_{cap} (a practical cap to control effort and avoid oversized CH sets when coverage is hard). K_{cap} is used only during construction and should not be confused with $K_{\max}^{\mathrm{route}}$, which is a deterministic per-round bound used solely to normalize the routing-cost term.

  At each step, GRASP ranks candidates by a greedy key, forms a restricted candidate list controlled by γ, samples one CH from that list, and continues [31] [32]. If enabled, Local-Search-Swap refines the seed by swapping a CH with a nearby non-CH node using a first-improvement rule, up to $L_{\max}$ accepted improvements while scanning at most k_{nn} neighbors per move. Each seed is evaluated through the deterministic pipeline (repair, assignment, routing) and may enter the elite pool $\mathcal{P}_{\mathrm{elite}}$, which keeps at most B seeds ranked by $\mathrm{Fitness}_t(\cdot)$; near-duplicates can be discarded using a Jaccard similarity threshold δ.

- **Phase II (fixed-set biased constructions):** From $\mathcal{P}_{\mathrm{elite}}$, we learn a fixed set $\mathcal{F}_t$ by counting how often each node appears in elite seeds. We retain nodes above frequency threshold τ and filter out low-energy nodes using threshold θ. We then rerun the same GRASP construction, starting from $\mathcal{H} \leftarrow \mathcal{F}_t$ when $\mathcal{F}_t \neq \emptyset$ (otherwise from $\emptyset$), with the same optional swap refinement and elite updates as in Phase I.

Finally, we select the best elite seed $\mathcal{H}_t^{\star}$ (minimum $\mathrm{Fitness}_t(\cdot)$) and deploy the repaired configuration $\mathcal{H}_t^{\star+} = \mathrm{Repair}_t(\mathcal{H}_t^{\star})$. Figure 2 presents an FSS-WSN round-level pipeline.

Algorithm 1 FSS-WSN (round t at the BS)

Input: Alive nodes V_t with their energies $E(t)$
Input: Parameters $(B, \gamma, K_{\text{cap}}, L_{\max}, k_{\text{nn}}, \tau, \theta, \delta, I_1, I_2)$
Output: Deployed CH set $\mathcal{H}_t^{\star+}$
$\mathcal{P}_{\text{elite}} \leftarrow \emptyset$
for $iter = 1$ **to** I_1 **do**
  $\mathcal{H} \leftarrow \text{GRASPCONSTRUCT}(V_t, E(t), \gamma, K_{\text{cap}})$ ▷ Phase I: populate the elite pool
  if $L_{\max} > 0$ **then**
    $\mathcal{H} \leftarrow \text{LOCALSEARCHSWAP}(\mathcal{H}, L_{\max}, k_{\text{nn}})$
  end if
  $\text{UPDATEELITE}(\mathcal{P}_{\text{elite}}, \mathcal{H}, B, \delta)$
end for
$\mathcal{F}_t \leftarrow \text{LEARNFIXEDSET}(\mathcal{P}_{\text{elite}}, \tau, \theta)$
for $iter = 1$ **to** I_2 **do**
  $\mathcal{H} \leftarrow \mathcal{F}_t$ **if** $\mathcal{F}_t \neq \emptyset$ **else** $\emptyset$ ▷ Extract recurrent components
  $\mathcal{H} \leftarrow \text{GRASPCOMPLETE}(\mathcal{H}, V_t, E(t), \gamma, K_{\text{cap}})$ ▷ Phase II: construct with $\mathcal{F}_t$ as a bias
  if $L_{\max} > 0$ **then**
    $\mathcal{H} \leftarrow \text{LOCALSEARCHSWAP}(\mathcal{H}, L_{\max}, k_{\text{nn}})$
  end if
  $\text{UPDATEELITE}(\mathcal{P}_{\text{elite}}, \mathcal{H}, B, \delta)$
end for
$\mathcal{H}_t^{\star} \leftarrow \arg\min_{\mathcal{H} \in \mathcal{P}_{\text{elite}}} \text{Fitness}_t(\mathcal{H})$
return $\text{Repair}_t(\mathcal{H}_t^{\star})$

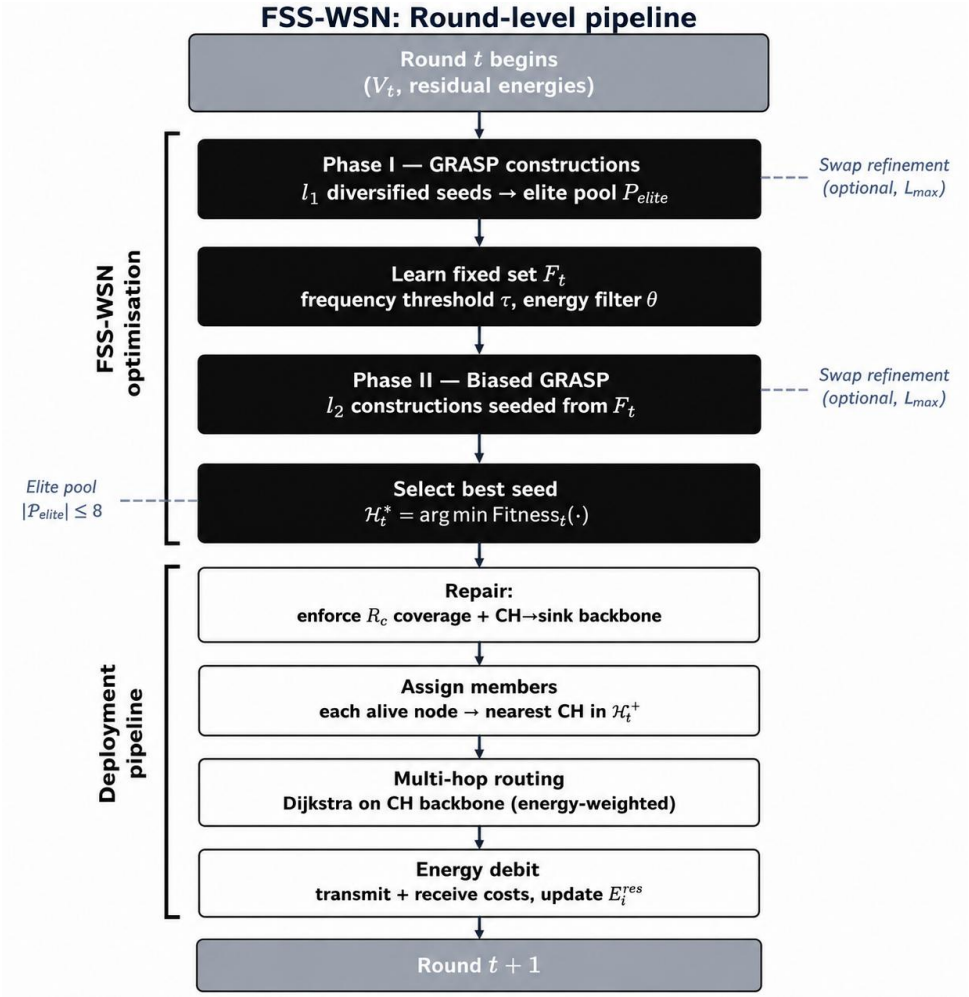


Figure 2. FSS-WSN round-level pipeline (black: BS-side optimization; white: deterministic deployment steps).

3.2 Computational Complexity

Let $N = |V_t|$ and $K = |\mathcal{H}^+|$. One fitness evaluation is dominated by nearest-CH assignment, $O(NK)$, and shortest-path computation on the CH backbone (Dijkstra), $O\big((K + |E^{mh}|)\log K\big)$, hence:

$$T_{\text{eval}} = O\big(NK + (K + |E^{mh}|)\log K\big). \tag{6}$$

Across a round, we perform $I_1 + I_2$ constructions. Because construction length is capped by K_{cap} and swap refinement (when used) is bounded by $L_{\max}$ accepted moves (each scanning k_{nn} candidates), the number of evaluations per construction is bounded by a constant $C_{\text{LS}}(K_{\text{cap}}, L_{\max}, k_{\text{nn}})$. Therefore,

$$T_{\text{round}} = O\big((I_1 + I_2)\, C_{\text{LS}}\, T_{\text{eval}}\big), \tag{7}$$

up to lower-order costs for elite maintenance and fixed-set learning. In the recommended configuration, $L_{\max} = 0$, so swap refinement is disabled and C_{LS} simplifies accordingly.

4. Experimental Setup and Configuration

All methods are evaluated within the same simulator core and under the same execution rules: strict clustering feasibility under R_c, routing on the CH-induced backbone, and the deterministic feasibility mapping $\text{Repair}_t(\cdot)$. Each run is simulated for up to $R_{\max} = 2500$ rounds, or until no node remains alive.

4.1 Compared Methods and Literature Positioning

FSS-WSN is compared against 11 baselines under identical conditions:

- **Metaheuristic optimizers (centralized):** PSO [12], GWO [13], ABC [14], SO [15], GJO [16], EMO-GJO [20], ESO–GJO [21].
- **Distributed protocols:** LEACH [6], HEED [7], SEP [8], EEM-LEACH-ABC [10].

Centralized optimizers (including FSS-WSN) run at the BS, which knows node positions *a priori* and collects one residual-energy scalar ($l_{\text{ctrl}} = 200$ bits) per alive node per round. The resulting control-plane overhead is $\approx$ 1.25 mJ/round (centre BS) to $\approx$ 1.98 mJ/round (corner BS), i.e. $< 4\,\%$ to $< 7\,\%$ of the per-round energy budget ($\approx$ 31.7 mJ), borne identically by all eight centralized methods.

Table 2 confirms that all methods share the standard LEACH radio constants; EEM-LEACH-ABC originally uses different parameters and was re-implemented under ours.

Table 2. Simulation parameters *vs.* literature (≡: identical to the previous).

Method	Field (m)	N	E_0 (J)	E_{elec}	l (bits)	BS position
LEACH	100^2	100	0.5	50	4000	(50,175)
HEED	100^2	100	0.5	≡	≡	center
SEP	100^2	100	0.5	≡	≡	(50,50)
PSO-CH	100^2	100	0.5	≡	≡	center
GWO-CH	100^2	100	0.5	≡	≡	center
ABC-CH	100^2	100	0.5	≡	≡	center
SO-CH	100^2	100	0.5	≡	≡	center
GJO-CH	100^2	100	0.5	≡	≡	center
EMO-GJO	100^2	100	0.5	≡	≡	center
ESO–GJO	100^2	100	0.5	≡	≡	center
EEM-LEACH-ABC	250^2	150	0.5	55	4400	(100,250)
This work	100^2	100	0.5	50	4000	center & corner

Common constants shared by all rows except EEM-ABC: $\epsilon_{fs} = 10$ pJ/bit/m^2, $\epsilon_{mp} = 0.0013$ pJ/bit/m^4, $E_{DA} = 5$ nJ/bit, $d_0 = 87$ m, $l_{\text{ctrl}} = 200$ bits.

4.2 Simulation Platform and Energy Model

The evaluation platform is a custom Python 3.12.3 simulator (NumPy 1.26.4) implementing the first-order radio energy model. All twelve methods execute within the same simulation core—identical energy accounting, Dijkstra-based multi-hop routing, and packet-level forwarding—so that differences in delivered utility reflect differences in the CH-selection strategy only.

The simulator operates at the network layer; MAC-layer contention and PHY-layer impairments are abstracted out, in line with all the referenced baselines. Experiments run on Windows 10 Pro, AMD Ryzen 5 7600X (6C/12T), 32 GB RAM. Source code: `https://github.com/RaoufOuanis/wsn-fss-simulation`.

Energy model: Fixed packet sizes: $l = 4000$ bits, $l_{\text{ctrl}} = 200$ bits; $E_{\text{elec}} = 50$ nJ/bit, $\epsilon_{fs} = 10$ pJ/bit/m^2, $\epsilon_{mp} = 0.0013$ pJ/bit/m^4, $d_0 = 87$ m, $E_{DA} = 5$ nJ/bit. Member-to-CH links are single-hop under R_c; backbone forwarding uses Dijkstra under r_{tx}. A hop the endpoint of which lacks energy is silently dropped.

Topology: $N = 100$ nodes, uniform 100×100 m field, $R_c = 25$ m. Two energy profiles and two BS positions (center/corner) yield the four configurations of Table 3.

Table 3. Network scenarios and topological parameters.

Parameter	S1 (Homogeneous)	S2 (Heterogeneous)
Field ($L \times L$)	100×100 m	100×100 m
N	100	100
Energy (J)	$E_0 = 0.5$	$E_0 = 0.5$ $E_{\text{adv}} = 1.0$
Percentage of advanced nodes (m)	–	20%
BS position	center / corner	center / corner
R_{max}	2500	2500

4.3 Algorithm Configuration

All iterative optimizers (FSS/PSO/GWO/ABC/SO/GJO/EMO-GJO/ESO–GJO) use the same per-round iteration budget, $n_{\text{iter}} = 60$. For population-based methods, we use pop_size$= 30$ as a common setting. For FSS-WSN, the budget is split as implemented: $I_1 = \text{round}(0.60\, n_{\text{iter}})$ and $I_2 = n_{\text{iter}} - I_1$. All hyper-parameters are summarized in Table 4**.**

Table 4. Algorithm configuration and hyper-parameters.

Algorithm	Parameter	Value
FSS-WSN	Iterations ($I_1 + I_2$)	$36 + 24 = 60$
	Elite pool size (B)	10
	RCL parameter (γ)	0.2
	Fixed-set thresholds (τ, θ)	0.6, 0.3
	Fitness weights ($w_1, w_2, w_3, w_R, \lambda$)	0.4, 0.4, 0.2, 0.05, 1.0
	Greedy-score weights ($\alpha_1, \alpha_2, \alpha_3$)	0.5, 0.3, 0.2
	Construction cap (K_{cap}) *	20
	Swap refinement ($L_{\max}, k_{nn}$)	disabled (0, 0)
All swarm methods below: $n_{iter} = 60$ *and pop* = 30*, k ∈ [1, 20]		
PSO-WSN	Inertia ω; c_1, c_2	0.7; 1.5, 1.5
GWO-WSN	Control a	$2 \rightarrow 0$ (Linear decay)
ABC-WSN	Limit; pop	10; 20**
SO, GJO, EMO-GJO, ESO-GJO	Default settings from original papers	
LEACH	p_{opt}	0.05
HEED	p_{init}, $c_{\min}$, n_{iter}	0.05, 0.02, 3
SEP	p_{opt}, m, E_0, E_{adv}	0.05, 0.2, 0.5, 1.0
EEM-LEACH-ABC	c_r, μ, epochs	0.10, 0.70, 5

* K_{cap} is a GRASP construction cap, whereas $K_{\max}^{\text{route}}$ is a deterministic, per-round normalization bound used only inside the routing-cost term and is not a tunable hyperparameter. ** ABC uses $pop = 20$ food sources (2× evaluations/iteration).

4.4 Evaluation Protocol

Every method produces a seed CH set $\mathcal{H}_0$, which may be empty or infeasible. The deterministic mapping $\mathcal{H}^+ = \text{Repair}_t(\mathcal{H}_0)$ (Eq. (4)) is applied before any routing or energy debiting, enforcing a non-empty CH set, strict R_c coverage, and CH-to-sink reachability under r_{tx}. Members attach to their nearest CH (smallest-index tie-breaking); routing follows Dijkstra on the induced backbone (Sub-section 2.2).

Protocol baselines (LEACH, HEED, SEP, EEM-LEACH-ABC) invoke repair once per round. Iterative optimizers invoke it at every fitness evaluation (Eq. (5)). The average number of CHs added by repair is reported for transparency.

Monte-Carlo protocol: Each configuration is evaluated over 30 paired seeds. A given seed fixes the node deployment—and, in S2, the advanced-node identities—across all methods; per-round randomness derives deterministically from the seed and round index. Results: $\mu \pm \sigma$; paired Wilcoxon signed-rank tests ($\alpha = 0.05$) with Holm correction.

Reported metrics: Wall-clock CPU time of the CH-selection routine (single machine); NFE for iterative methods only. Lifetime markers: FND, HND, LND, R_{last}. Primary endpoint: cumulative delivered throughput.

Ablation and sensitivity: Ablation results ($\pm$Phase II, $\pm$regularizer) and one-at-a-time sensitivity sweeps over $(\gamma, \tau, \theta, n_{\text{iter}})$ are reported in Sub-section 5.4 following standard metaheuristic validation practice [33] [34] [35] [36].

5. EXPERIMENTAL RESULTS

We evaluate FSS-WSN under the protocol described in Section 4: common simulator core, first-order energy accounting, strict clustering under R_c, multi-hop CH to sink forwarding, and deterministic hard-feasibility enforcement *via* $\text{Repair}_t(\cdot)$. All results are reported as mean $\mu \pm \sigma$ over 30 paired Monte-Carlo seeds. Paired Wilcoxon signed-rank tests are computed on matched seeds. Since multiple baselines are compared to FSS-WSN within a configuration, the Holm correction is applied within that configuration.

5.1 Protocol Recap and Reported Endpoints

All methods run for up to $R_{\max} = 2500$ rounds in a centralized and round-based mode. For each seed, the same node deployment (and, in S2, the same advanced-node identities) is reused across methods. All methods -including classical protocols- are evaluated under the same convention of feasibility: the runner applies $\text{Repair}_t(\cdot)$ before routing and energy accounting, enforcing strict R_c coverage and CH→sink reachability under r_{tx}. We report:

- **Primary endpoint:** is the cumulative useful delivered reports, denoted **Throughput**.
- **Service duration:** R_{last} and FND/HND/LND.
- **Traffic proxy:** CH→sink (pkts).
- **Compute cost:** per-round CPU time as the main indicator; NFE only as an implementation-dependent proxy.

5.2 Primary Endpoint: Delivered Utility across Configurations

Table 5 summarizes **Throughput (reports)** across the four configurations (S1/S2 with centered/corner BS). The pattern is stable: **FSS-WSN achieves the highest delivered utility in all configurations.** A visual summary against the best baseline in each scenario is provided in Figure 3. Concretely:

- **S1 (homogeneous):** $119{,}533 \pm 1{,}190$ (center) and $97{,}064 \pm 1{,}610$ (corner), i.e., $+4.57\%$ *vs.* HEED (center) and $+3.16\%$ *vs.* GWO (corner).
- **S2 (heterogeneous):** $140{,}131 \pm 5{,}210$ (center) and $116{,}025 \pm 3{,}829$ (corner), i.e., $+3.68\%$ *vs.* HEED (center) and $+2.40\%$ *vs.* GWO (corner).

Positioning *vs.* SO/GJO-family variants (EMO–GJO, ESO–GJO): Under the unified evaluation protocol used in this study, FSS-WSN outperformed both EMO-GJO and ESO-GJO in all four configurations. In terms of delivered reports, FSS-WSN achieved gains of 8.17% in S1-center, 3.51%

in S1-corner, 5.88% in S2-center, and 5.08% in S2-corner relative to the best SO/GJO-family variant. These results suggest that part of the advantage reported for EMO-GJO and ESO-GJO in prior studies may depend on differences in feasibility handling and routing assumptions across simulation frameworks.

Table 5. Primary endpoint (delivered reports) across configurations (mean ± std, 30 seeds).

Method	S1-center	S1-corner	S2-center	S2-corner
FSS-WSN	119,533 ± 1,190	97,064 ± 1,610	140,131 ± 5,210	116,025 ± 3,829
ESO–GJO	110,389 ± 1,581	93,004 ± 1,576	132,312 ± 4,954	110,412 ± 3,332
EMO–GJO	110,377 ± 1,595	93,774 ± 1,591	132,352 ± 4,988	108,218 ± 3,670
GJO	110,388 ± 1,608	91,525 ± 1,559	132,313 ± 5,015	108,168 ± 3,205
SO	110,500 ± 1,603	91,398 ± 1,511	132,294 ± 5,010	107,508 ± 3,026
PSO	110,433 ± 1,626	93,484 ± 1,558	132,019 ± 4,967	109,888 ± 3,431
GWO	110,417 ± 1,596	94,092 ± 1,335	131,991 ± 4,973	113,309 ± 3,597
ABC	110,372 ± 1,622	92,623 ± 1,543	132,423 ± 4,988	110,144 ± 3,504
HEED	114,306 ± 1,734	87,358 ± 1,366	135,153 ± 5,676	107,175 ± 3,488
LEACH	110,276 ± 1,624	87,714 ± 1,546	131,205 ± 4,976	105,366 ± 2,610
SEP	107,190 ± 1,289	91,247 ± 1,487	131,550 ± 4,453	107,612 ± 3,422
EEM-LEACH-ABC	109,658 ± 895	90,471 ± 1,603	131,421 ± 4,605	109,348 ± 3,999

All 44 FSS-WSN *vs.* baseline differences significant (paired Wilcoxon, $p_{Holm} < 0.001$, $r_{rb} = 1.0$).

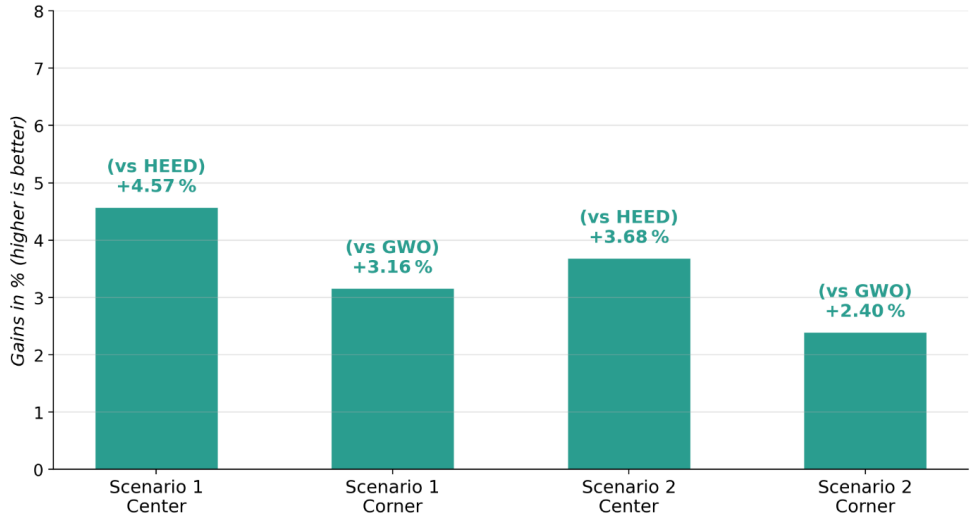


Figure 3. Primary endpoint summary (delivered reports): FSS-WSN compared to best baseline.

Positioning of EEM-LEACH-ABC: We added EEM-LEACH-ABC, as it is the most recent protocol in this problem family. Under our unified evaluation protocol, it ranked in the lower half across all configurations (11th in S1-center and S2-center, 10th in S1-corner, and 6th at best in S2-corner). Its throughput remained 6.11% to 9.00% below that of FSS-WSN, suggesting that ABC-based parameter tuning within a LEACH-style framework is not sufficient to match direct CH-set optimization in the present setting.

5.3 Service Duration, Robustness, and Computational Cost

Advantage persists under routing-coupled multi-hop: Two observations support the view that the throughput advantage is not merely an artifact of the experimental setup:

(i) Lower repair dependence: $\text{Repair}_t(\cdot)$ silently adds CHs or relays to force feasibility, so any method that relies heavily on repair is partly "cheating"—its throughput is artificially propped up by nodes that it did not choose. FSS-WSN barely triggers repair at all: 0.27 added CH/round in S1-center, as low as 0.004 in S1-corner, and similar numbers in S2. The EMO/ESO family, by contrast, needs ≈ 2.6–3.5

insertions per round. This aligns perfectly with the regularizer's goal: the optimizer has already learned to produce near-feasible solutions on its own.

(ii) Lower relay hotspots in corner geometry: Corner placement stretches multi-hop paths, and a few relay nodes inevitably become bottlenecks. What we found is that FSS-WSN nearly halves the worst-case relay load: mh_q_max drops from $\approx$ 22.4 to 12.83 (S1-corner) and 14.01 (S2-corner). Concretely, fewer packets pile up at any single relay, so fewer get dropped along the chain.

Service duration and tail behavior: Table 6 lists R_{last} for all methods. One thing that we noticed right away is that a longer R_{last} can be misleading: several baselines linger for dozens of extra rounds while barely delivering anything. Figure 4. Cumulative delivered reports over rounds (mean, 30 seeds). FSS-WSN plateaus highest in all scenarios depite shorter operational lifetime in some cases captures this well—FSS-WSN's curve climbs faster, finishes higher, and then the network dies. HEED and GWO approach FSS-WSN's final level in certain setups, but do not reach it; LEACH and SEP lag by a wider margin.

Table 6. Operational lifetime R_{last} across configurations (mean $\pm$ std, 30 paired seeds).

Method	S1-center	S1-corner	S2-center	S2-corner
FSS-WSN	1575 ± 56	1400 ± 78	2349 ± 51	1560 ± 62
ESO–GJO	$1674 \pm 50^{\ddagger}$	$1620 \pm 68^{\ddagger}$	$2419 \pm 111^{\ddagger}$	$1781 \pm 204^{\ddagger}$
EMO–GJO	$1684 \pm 54^{\dagger}$	$1590 \pm 65^{\ddagger}$	$2440 \pm 105^{\ddagger}$	$1772 \pm 219^{\ddagger}$
GJO	$1685 \pm 52^{*}$	$1612 \pm 65^{\ddagger}$	$2417 \pm 100^{\ddagger}$	$1773 \pm 219^{\ddagger}$
SO	$1684 \pm 48^{\dagger}$	$1598 \pm 65^{\ddagger}$	$2440 \pm 105^{\ddagger}$	$1773 \pm 219^{\ddagger}$
PSO	$1692 \pm 54^{\ddagger}$	$1612 \pm 66^{\ddagger}$	$2438 \pm 104^{\text{ns}}$	$1773 \pm 219^{\ddagger}$
GWO	$1690 \pm 52^{\dagger}$	$1490 \pm 61^{\ddagger}$	$2438 \pm 102^{\text{ns}}$	$1773 \pm 219^{\ddagger}$
ABC	$1692 \pm 53^{\text{ns}}$	$1612 \pm 67^{\ddagger}$	$2440 \pm 105^{\dagger}$	$1773 \pm 219^{\ddagger}$
HEED	$1812 \pm 59^{\ddagger}$	$1529 \pm 57^{\text{ns}}$	$2294 \pm 111^{*}$	$1605 \pm 218^{\ddagger}$
LEACH	$1724 \pm 48^{\ddagger}$	$1524 \pm 44^{\ddagger}$	$2419 \pm 112^{\dagger}$	$1802 \pm 240^{\ddagger}$
SEP	$1513 \pm 70^{\ddagger}$	$1603 \pm 58^{\ddagger}$	$2440 \pm 105^{\dagger}$	$1773 \pm 219^{\ddagger}$
EEM-LEACH-ABC	$1797 \pm 120^{\dagger}$	$1527 \pm 122^{\ddagger}$	$2499 \pm 0^{\dagger}$	$2358 \pm 195^{\ddagger}$

Significance markers are based on paired Wilcoxon signed-rank tests (Holm-corrected): ‡ $p_{\text{Holm}} < 0.001$; † $p_{\text{Holm}} < 0.01$; $*$ $p_{\text{Holm}} < 0.05$; $^{\text{ns}}$ not significant. A positive rrb indicates that FSS-WSN has a longer R_{last}.

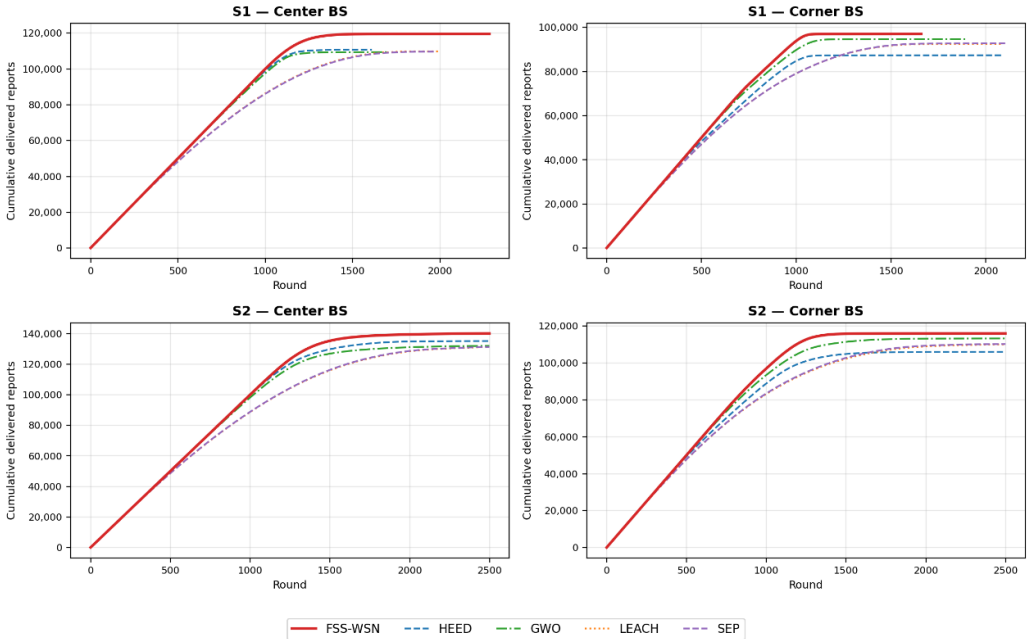


Figure 4. Cumulative delivered reports over rounds (mean, 30 seeds). FSS-WSN plateaus highest in all scenarios depite shorter operational lifetime in some cases.

Wilcoxon tests back this up: in corner configurations, FSS-WSN terminates significantly earlier than 10 of 11 baselines ($p_{\text{Holm}} < 0.001$). HEED is the sole exception—its aggressive clustering sometimes drains the network even faster than FSS-WSN does. The pattern is consistent: FSS-WSN pushes harder per round, finishes sooner, but accumulates more reports overall.

Practical significance of the throughput gain: The observed throughput improvement (2.4–4.6%) corresponds to approximately 2,700–5,200 additional reports delivered to the BS, which is equivalent to about 27–52 rounds of full service in a 100-node network. When combined with the 17–22× reduction in per-round CPU time relative to the fastest swarm baseline (Table 7), these results indicate that FSS-WSN offers a particularly favorable throughput–computation trade-off in the evaluated settings (Figure 5). This interpretation is supported by the statistical analysis, as all 44 pairwise throughput comparisons against FSS-WSN were significant ($r_{rb} = 1.0$, $p_{\text{Holm}} < 0.001$).

Computational cost: Table 7 and Figure 5 complete the picture: FSS-WSN is the cheapest iterative optimizer and the highest in delivered utility, in every configuration. The 17–22× CP advantage over the fastest swarm baseline stems in part from fitness memoization; the reported NFE counts include cache hits, so effective unique evaluations are substantially fewer. All 28 pairwise CPU-time comparisons are significant ($p_{\text{Holm}} < 0.001$, $r_{rb} = 1.0$). Protocol baselines (LEACH, HEED, SEP, EEM-LEACH-ABC) are lightweight by design and are not reported.

Table 7. Per-round computational overhead for iterative optimizers (CPU time in seconds).

Scenario	Method	CPU (s)	NFE
S1-center	FSS-WSN	**0.085**	**600 ± 91**
	SO	1.440 ± 0.047	1830 ± 0
	GJO	1.448 ± 0.047	1830 ± 0
	EMO–GJO	1.451 ± 0.053	1800 ± 0
	ESO–GJO	1.428 ± 0.077	1800 ± 0
	PSO	1.431 ± 0.077	1800 ± 0
	GWO	1.425 ± 0.067	1800 ± 0
	ABC	2.068 ± 0.087	2528 ± 4
S1-corner	FSS-WSN	**0.076**	**754 ± 107**
	SO	1.357 ± 0.032	1830 ± 0
	GJO	1.362 ± 0.032	1830 ± 0
	EMO–GJO	1.370 ± 0.026	1800 ± 0
	ESO–GJO	1.450 ± 0.033	1800 ± 0
	PSO	1.412 ± 0.039	1800 ± 0
	GWO	1.802 ± 0.143	1800 ± 0
	ABC	2.198 ± 0.261	2527 ± 5
S2-center	FSS-WSN	**0.0885**	**711 ± 89**
	SO	1.982 ± 0.072	1830 ± 0
	GJO	2.159 ± 0.071	1830 ± 0
	EMO–GJO	3.029 ± 0.074	1800 ± 0
	ESO–GJO	2.981 ± 0.052	1800 ± 0
	PSO	2.050 ± 0.056	1800 ± 0
	GWO	2.088 ± 0.070	1800 ± 0
	ABC	3.152 ± 0.091	2529 ± 5
S2-corner	FSS-WSN	**0.067**	**860 ± 124**
	SO	1.112 ± 0.180	1830 ± 0
	GJO	1.190 ± 0.267	1830 ± 0
	EMO–GJO	1.103 ± 0.209	1800 ± 0
	ESO–GJO	1.323 ± 0.276	1800 ± 0
	PSO	1.370 ± 0.271	1800 ± 0
	GWO	1.215 ± 0.282	1800 ± 0
	ABC	1.695 ± 0.367	2531 ± 6

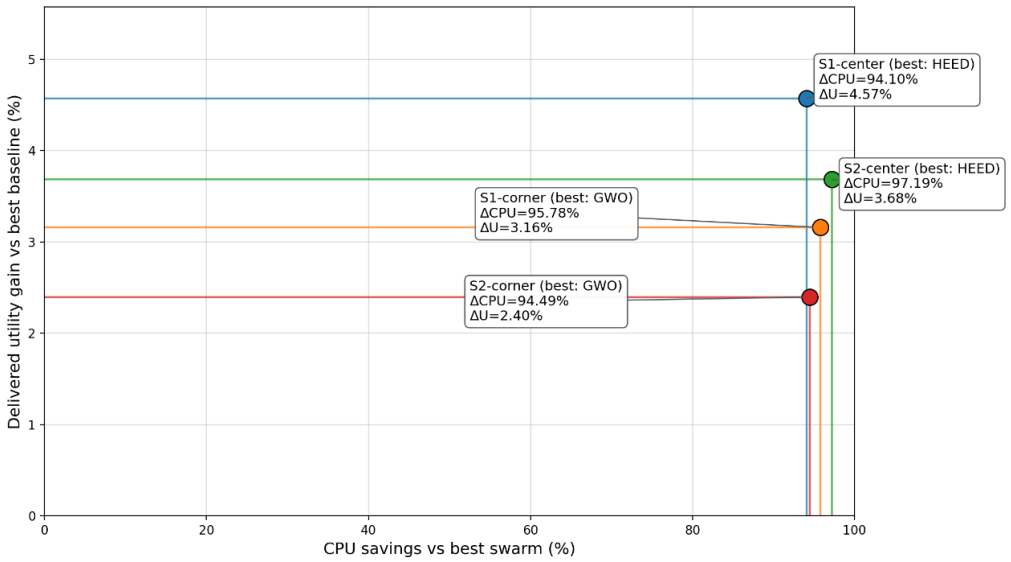


Figure 5. CPU–utility trade-off of FSS-WSN *vs.* the strongest iterative competitor per scenario.

5.4 Ablation Study

To quantify the contribution of the two main components added on top of Phase I GRASP, we run 30-seed ablation experiments on the four test settings. Two algorithmic variants are considered: (1) without Phase II, which disables fixed-set intensification and relies on Phase I alone; and (2) without regularizer, which sets $\lambda = 0$ and removes the repair-dependence penalty from the objective. All other parameters remain as in Table 4, and the same paired seeds are used throughout. Table 8 reports relative changes with respect to the full FSS-WSN configuration. For reference, on S1-center (the setting used for detailed reporting), the absolute throughput values are: $119{,}533 \pm 1{,}190$ (full FSS-WSN), $118{,}534 \pm 1{,}190$ (without Phase II), and $118{,}859 \pm 1{,}198$ (without regularizer).

Table 8. Ablation study: relative changes with respect to full FSS-WSN (30 paired seeds).

	S1-center	**S1-corner**	**S2-center**	**S2-corner**
Δ *Throughput (%)*				
without Phase II	-0.84^{ns}	$-1.18^{\ddagger}$	-1.01^{ns}	-1.07^{*}
without regularizer	$-0.56^{\ddagger}$	-0.80^{ns}	$-0.20^{\ddagger}$	-1.02^{ns}
Δ *Per-round CPU time (%)*				
without Phase II	-36	-42	-33	-41
without regularizer	-5	$+1$	-1	-3

Table 8 reports throughput changes in % and CPU changes as relative per-round time; statistical significance is assessed *via* paired Wilcoxon tests ($\ddagger$ $p < 0.001$; $*$ $p < 0.05$; $^{\text{ns}}$ not significant). Removing Phase II consistently lowers mean throughput across the four test settings, although the effect remains modest in magnitude. This indicates that, at $N = 100$, the marginal contribution of Phase II over a Phase-I-only variant is small, but consistent. We therefore keep Phase II in the reference configuration: it delivers the best mean throughput in all cases and provides a principled intensification mechanism, at the cost of roughly one-third of the per-round CPU time (33%–42% reduction when disabled). Removing the repair-dependence regularizer also leads to small, but consistently negative, throughput changes, with negligible runtime impact ($|\Delta| \leq 5\%$). We retain it, because it biases the search away from solutions the quality of which depends heavily on repair; in that sense, the regularizer acts primarily as a robustness-oriented term rather than as a direct throughput booster.

One-at-a-time sensitivity sweeps: Figure 6 reports the throughput variation when each of the four main hyper-parameters (γ, τ, θ, n_{iter}) is varied individually while the others are kept at the defaults of Table 4. Each bar shows the percentage throughput change relative to the default value (marked with $\star$), averaged over 5 paired seeds on S1-center (2 500 rounds). All variations remain below $\pm 1\%$ in throughput (τ, θ, and n_{iter} stay within $\pm 0.05\%$; only γ reaches $\approx 0.9\%$), confirming that the selected

configuration is not a local optimum artifact and that FSS-WSN performance is robust to moderate hyperparameter perturbations.

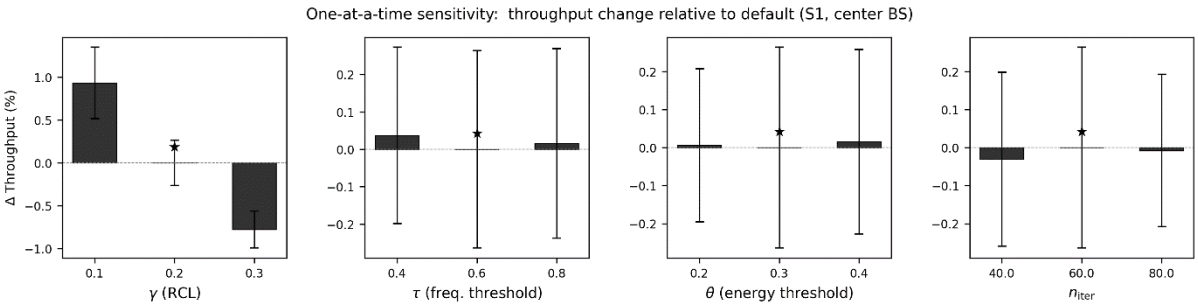


Figure 6. One-at-a-time sensitivity: throughput change relative to default (S1, center BS, 5 seeds, 2 500 rounds). All deviations are below 1%; the ⋆ marks the retained configuration.

5.5 Scalability $N = 200$

To assess how FSS-WSN scales beyond the 100-node configuration, we ran experiments with $N = 200$ homogeneous nodes in the same 100×100 m area (S1, center BS, $E_0 = 0.5$ J, 10 seeds, $R_{\max} = 2000$). Table 9 shows that FSS-WSN maintains the highest throughput (238,707, +8.7% *vs.* HEED) and the most stable FND (−2.8% degradation *vs.* −15% to −20% for LEACH/SEP). Per-round CPU time scales from ∼0.085 s (N =100) to ∼0.20 s (N =200), consistent with the complexity of Eq. (7), all FSS-WSN *vs.* baseline differences are significant (paired Wilcoxon, $p = 0.002$, $r_{rb} = 1.0$).

Table 9. Scalability: N =100 (30 seeds) *vs.* N =200 (10 seeds), S1 center BS.

	Throughput		**FND (rounds)**	
	$N = 100$	$N = 200$	$N = 100$	$N = 200$
FSS-WSN	119,533	238,707	975	947 (−2.8%)
HEED	114,306	219,706	812	805 (−0.9%)
LEACH	110,276	216,869	201	172 (−14.7%)
SEP	107,190	216,413	200	161 (−19.7%)

5.6 Scope and Threats to Validity

These results concern a configuration that assumes a centralized round-based decision, with multi-hop CH to sink routing, under a strict deterministic feasibility convention for all methods. Interpretation is delimited by:

- **Routing abstraction and MAC/PHY gap:** The simulator operates at the network layer: Dijkstra provides optimal multi-hop relay paths, but MAC-layer contention (e.g., CSMA/CA back-off, duty cycling) and PHY-layer effects (fading, interference, packet-error rate) are abstracted out. This is the standard evaluation framework for all twelve compared methods.

  Three arguments bound the impact of this abstraction on the *relative* comparison (FSS-WSN *vs.* baselines): (i) MAC losses are topology-dependent and round-dependent, but all methods produce comparable cluster counts (6–12 CHs/round) and comparable traffic patterns, so that the expected MAC loss ratio is similar across methods; (ii) the FSS-WSN advantage stems from *better CH placement* (lower relay hotspots, lower repair dependence), which reduces spatial congestion—a feature that would *improve* rather than degrade under a contention-based MAC; (iii) prior comparative studies using full-stack simulators for LEACH-family protocols report that relative algorithm rankings are preserved even though absolute packet-delivery ratios decrease . Nonetheless, full-stack validation (ns-3 or Cooja with IEEE 802.15.4 MAC) remains an explicit direction for future work.

- **Feasibility convention:** Deterministic repair enforces strict radius coverage and backbone reachability; alternative constraint handling may change absolute metric scales.

- **Bounded surrogate:** Normalization makes the optimization more stable across rounds under a fixed budget, but it can also “compress” differences among weak candidates. Therefore, our conclusions rely primarily on end-to-end system metrics, notably Throughput.
- **Central observability:** The BS is assumed to know all node positions (fixed, from pre-deployment survey) and to receive a single residual-energy scalar from each alive node at every round. As quantified in Subsection 4.1, this control-plane overhead represents less than 4 % (centre BS) to 7 % (corner BS) of the total network energy per round, and is borne identically by all eight centralized optimizers.
- **Compute-cost dependence:** CPU time is platform-dependent; we report it on a fixed platform and complement it with NFE as a secondary proxy.

Within this scope and the paired Monte-Carlo protocol, FSS-WSN shows a consistent delivered-utility advantage across the four configurations.

6. CONCLUSION

This work presented FSS-WSN and demonstrated its usefulness for Wireless Sensor Networks. By adding a learning mechanism to the well-known GRASP metaheuristic, the proposed approach naturally fits a centralized decision-making setting for selecting cluster heads. It improves overall network utility through a more guided search that is faster and less costly. We enforced strict feasibility through a deterministic repair procedure, since our goal was to implement a realistic and deployable approach.

The simulations, which covered both favorable and unfavorable configurations, and included diverse and representative baselines (including classical protocols, population-based metaheuristics, and more recent optimizers), demonstrated across all settings (four scenarios) that FSS-WSN consistently achieved the highest “cumulative throughput” until the network becomes inactive, with gains ranging from 2.4% to 4.57% over the best baseline. BS-side CPU time was dramatically lower than that of the strongest iterative competitors, with reductions ranging from 94.1% to 97.19% depending on the scenarios.

The results also indicate that the approach is particularly effective in scenarios with a high risk of bottlenecks. However, compared to some protocol baselines, FSS-WSN can be less favorable in terms of LND (Last Node Die). Overall, this positions FSS-WSN as a strong choice for use cases where the priority is to deliver as much useful information as possible, with a quick decision time (e.g., industrial monitoring, emergency response, sensitive perimeter surveillance, …etc.), and where the survival of the very last node is less meaningful if it does not correspond to end-to-end delivery.

Future work will focus on bringing the simulation even closer to real conditions by incorporating MAC/PHY constraints that are currently abstracted. Additionally, we will seek to better understand the utility–cost trade-off by using a fixed time limit per round when the BS has to decide under real deadlines.

REFERENCES

[1] J. N. Al-Karaki and A. E. Kamal, "Routing Techniques in Wireless Sensor Networks: A Survey," IEEE Wireless Communications, vol. 11, no. 6, pp. 6–28, 2004.

[2] K. Guleria and A. K. Verma, "Comprehensive Review for Energy Efficient Hierarchical Routing Protocols on Wireless Sensor Networks," Wireless Network, vol. 25, no. 4, pp. 1159–1183, 2019.

[3] C. Nakas, D. Kandris and G. Visvardis, "Energy Efficient Routing in Wireless Sensor Networks: A Comprehensive Survey," Algorithms, vol. 13, no. 3, p. 72, 2020.

[4] H. B. Salameh, M. Dhainat and E. Benkhelifa, "A Survey on Wireless Sensor Network-based IoT Designs for Gas Leakage Detection and Fire-Fighting Applications," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 5, no. 2, pp. 60–72, 2019.

[5] L. Chhaya et al., "Wireless Sensor Network Based Smart Grid Communications: Cyber Attacks, Intrusion Detection System and Topology Control," Electronics, vol. 6, no. 1, p. 5, 2017.

[6] W. R. Heinzelman, A. Chandrakasan and H. Balakrishnan, "Energy-efficient Communication Protocol for Wireless Microsensor Networks," Proc. of the 33rd Annual Hawaii Int. Conf. on System Sciences (HICSS), DOI: 10.1109/HICSS.2000.926982, Maui, HI, USA, 2000.

[7] O. Younis and S. Fahmy, "HEED: A Hybrid, Energy-efficient, Distributed Clustering Approach for Ad

Hoc Sensor Networks," IEEE Transactions on Mobile Computing, vol. 3, no. 4, pp. 366–379, 2004.
[8] G. Smaragdakis, I. Matta and A. Bestavros, "SEP: A Stable Election Protocol for Clustered Heterogeneous Wireless Sensor Networks," Proc. of the 2nd Int. Workshop on Sensor and Actor Network Protocols and Applications (SANPA), 2004.
[9] S. Arjunan and S. Pothula, "A Survey on Unequal Clustering Protocols in Wireless Sensor Networks," Journal of King Saud University - Computer and Information Sciences, vol. 31, no. 3, pp. 304–317, 2019.
[10] S. Zhang, X. Liu and M. Trik, "Energy Efficient Multi Hop Clustering Using Artificial Bee Colony Metaheuristic in WSN," Scientific Reports, vol. 15, p. 26803, 2025.
[11] R. Sharma, V. Vashisht and U. Singh, "Metaheuristics-based Energy Efficient Clustering in WSNs: Challenges and Research Contributions," IET Wireless Sensor Systems, vol. 10, no. 5, pp. 253–264, 2020.
[12] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," Proc. of Int. Conf. on Neural Networks (ICNN'95), DOI: 10.1109/ICNN.1995.488968, Perth, WA, Australia, 1995.
[13] S. Mirjalili, S. M. Mirjalili and A. Lewis, "Grey Wolf Optimizer," Advances in Engineering Software, vol. 69, pp. 46–61, 2014.
[14] D. Karaboga, "An Idea Based on Honey Bee Swarm for Numerical Optimization," [Online], Available: https://abc.erciyes.edu.tr/pub/tr06_2005.pdf, 2005.
[15] F. A. Hashim et al., "A Novel Meta-heuristic Optimization Algorithm Inspired by Snake Movement Patterns," Knowledge-based Systems, vol. 242, p. 108320, 2022.
[16] N. Chopra and M. M. Ansari, "Golden Jackal Optimization: A Novel Nature-Inspired Optimizer for Engineering Applications," Expert Systems with Applications, vol. 198, p. 116924, 2022.
[17] N. Gupta, A. B. b. A. Hamid, A. B. B. Mahat and A. Kumar, "Machine-learning-enhanced Glowworm Swarm Optimization for Energy-efficient Multi-hop Routing in Wireless Sensor Networks," Results in Control and Optimization, vol. 22, p. 100667, 2026.
[18] N. Gupta et al., "Analysis of Energy-efficient Smart Path Optimization Routing Protocol for Wireless Sensor Networks," Results in Engineering, vol. 28, p. 107456, 2025.
[19] C. A. C. Coello, "Theoretical and Numerical Constraint-handling Techniques Used with Evolutionary Algorithms: A Survey of the State of the Art," Computer Methods in Applied Mechanics and Engineering, vol. 191, no. 11–12, p. 1245–1287, 2002.
[20] T. Mazumder, B. V. R. Reddy and A. Payal, "Energy Based Multi Objective Golden Jackal Optimization for Cluster Based Routing in Wireless Sensor Network," Soft Computing, vol. 28, no. 20, pp. 11927–11943, 2024.
[21] Z. Wang, J. Duan and P. Xing, "Multi-hop Clustering and Routing Protocol Based on Enhanced Snake Optimizer and Golden Jackal Optimization in WSNs," Sensors, vol. 24, no. 4, p. 1348, 2024.
[22] S. Okdem and D. Karaboga, "Routing in Wireless Sensor Networks Using an Ant Colony Optimization (ACO) Router Chip," Sensors, vol. 9, no. 2, pp. 909–921, 2009.
[23] F. Glover, M. Laguna and R. Marti, "Principles and Strategies of Tabu Search," Handbook of Approximation Algorithms and Metaheuristics: Methodologies and Traditional Applications, 2nd Ed., T. F. Gonzalez, Ed., Chapman and Hall/CRC, pp. 573–597, 2018.
[24] R. Jovanovic, M. Tuba and S. Voss, "Fixed Set Search Applied to the Traveling Salesman Problem," Hybrid Metaheuristics, vol. 11380, pp. 63–77, Springer, 2019.
[25] R. Jovanovic and S. Voss, "Matheuristic Fixed Set Search Applied to the Multidimensional Knapsack Problem and the Knapsack Problem with Forfeit Sets," OR Spectrum, vol. 46, no. 4, pp. 1329–1365, 2024.
[26] T. S. Rappaport, Wireless Communications: Principles and Practice, 2nd Ed., Prentice Hall PTR, 2002.
[27] D. Ruan, J. Huang and X. Li, "Uneven Clustering Routing Algorithm Based on Energy and Distance for Wireless Sensor Networks," Journal of Sensors, vol. 2019, p. 8109616, 2019.
[28] M. R. Garey and D. S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W. H. Freeman and Company, 1979.
[29] S. Guha and S. Khuller, "Approximation Algorithms for Connected Dominating Sets," Algorithmica, vol. 20, no. 4, pp. 374–387, 1998.
[30] K. Deb, "An Efficient Constraint Handling Method for Genetic Algorithms," Computer Methods in Applied Mechanics and Engineering, vol. 186, no. 2-4, pp. 311–338, 2000.
[31] T. A. Feo and M. G. C. Resende, "Greedy Randomized Adaptive Search Procedures," Journal of Global Optimization, vol. 6, no. 2, pp. 109–133, 1995.
[32] M. G. C. Resende and C. C. Ribeiro, "Greedy Randomized Adaptive Search Procedures: Advances and Extensions," Handbook of Metaheuristics, Part of the Book Series: Int. Series in Operations Research & Management Science, vol. 272, pp. 169–220, Springer, 2018.
[33] H. J. C. Barbosa, H. S. Bernardino and A. M. S. Barreto, "Using Performance Profiles to Analyze the Results of the 2006 CEC Constrained Optimization Competition," Proc. of the IEEE Congress on Evolutionary Computation (CEC), DOI: 10.1109/CEC.2010.5586105, Barcelona, Spain, 2010.
[34] T. Bartz-Beielstein et al., Experimental Methods for the Analysis of Optimization Algorithms, ISBN: 978-3-642-02537-2, Berlin: Springer, 2010.

[35] T. Kadavy et al., "Impact of Boundary Control Methods on Bound-constrained Optimization Benchmarking," IEEE Transactions on Evolutionary Computation, vol. 26, no. 6, pp. 1207–1221, 2022.

[36] M. Lopez-Ibañez et al., "The Irace Package: Iterated Racing for Automatic Algorithm Configuration," Operations Research Perspectives, vol. 3, pp. 43–58, 2016.

ملخص البحث:

لا تـــزال شـــبكات الاستشـــعار اللاســـلكية مجـــالاً بحثيـــاً نشـــطاً فـــي كـــلٍّ مـــن القطـــاعين العســـكري والمـــدني، مدفوعـــة بتزايـــد تطبيقاتهـــا. وفـــي السّـــنوات الأخيـــرة، شـــهدنا تحـــوّلاً تـــدريجياً نحـــو دمـــج الـــذّكاء الاصـــطناعي لمواجهـــة التّحـــدّي المســـتمرّ المتمثّـــل فـــي تحســـين استهلاك الطّاقة في هذه الشّبكات.

فـــي هـــذا البحـــث، نقـــدّم تعـــديلاً جديـــداً لآليـــة البحـــث فـــي المجموعـــة الثّابتـــة لتُناســـب شـــبكات الاستشـــعار اللاســـلكية. وتضـــيف هـــذه الآليـــة مرحلـــة تعلُّـــم إلـــى خوارزميـــة (GRASP) المعروفـــة. ويعمـــل الـــنّهج المســـتخدم علـــى توجيـــه محطّـــة القاعـــدة فـــي بيئـــةٍ مركزيـــة متعـــدّدة القفـــزات لاختيـــار الـــرؤوس المثلـــى للمجموعـــات، الأمـــر الّـــذي يعظّـــم الفائـــدة الإجماليـــة للشبكة.

وقـــد قُمنـــا بتقيـــيم نهجنـــا فـــي ظـــلّ شـــروط عدالـــةٍ موثّقـــة، مقارنـــةً بمجموعـــة واســـعة مـــن المعـــايير الأساســـية المعتمـــدة، بمـــا فـــي ذلـــك بروتوكـــولات التّجميـــع اللاســـلكية، ومجسّـــات السّرب شائعة الاستخدام، والنّموذج الهجين متعدّد القفزات.

وأظهـــرت النّتـــائج تحسُّـــناً ذا دلالـــةٍ إحصـــائية مقارنـــةً بأفضـــل معيـــار أساســـي فيمـــا يتعلّـــق بعـــدد التّقـــارير المســـلّمة ووقـــت وحـــدة المعالجـــة المركزيـــة الـــلازم لاتّخـــاذ القـــرار. وتُشـــير هـــذه النّتـــائج إلـــى أنّ الـــنّهج الّـــذي اســـتخدمناه يمثّـــل خيـــاراً قويـــاً وعمليـــاً للعديـــد مـــن حـــالات استخدام شبكات الاستشعار اللاسلكية.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

ON THE EFFECT OF KEYHOLE CHANNEL IN RSMA NETWORKS: A THEORETICAL OUTAGE ANALYSIS

Hong-Nhu Nguyen[1] and Phong-Cuong Ngo[2]

(Received: 20-Feb.-2026, Revised: 16-Apr.-2026, Accepted: 30-Apr.-2026)

ABSTRACT

This paper examines the theoretical performance of rate-splitting multiple access (RSMA) in keyhole fading channels. This rank-deficient environment is notorious for degrading the performance of traditional single-input single-output (SISO) systems. For a two-user downlink channel, exact and asymptotic outage probability results of RSMA with perfect and imperfect successive interference cancellation (pSIC and ipSIC) are derived. Closed-form solutions are derived by comparing the product of two independent Nakagami-m fading channels that model the keyhole effect. To further demonstrate RSMA's robustness in such an environment, we also examine the diversity order and unveil the effect of keyhole-induced rank deficiency on system reliability. Our results demonstrate that RSMA retains a performance edge over non-orthogonal multiple access (NOMA), especially with imperfect SIC or low SNR. Numerical results and Monte-Carlo simulations confirm the theoretical formulae and show that RSMA can combat the harmful effects of the keyhole channel better than other conventional schemes. The results confirm the promise of RSMA for future wireless systems operating in severe-fading environments.

KEYWORDS

Keyhole channel, Rate-splitting multiple access (RSMA), Outage probability, Nakagami-m, Imperfect SIC, Diversity order, Non-orthogonal multiple access (NOMA).

1. INTRODUCTION

Rate-Splitting Multiple Access (RSMA) has been an effective non-orthogonal transmission technique that is capable of bridging and surpassing traditional methods like Space-Division Multiple Access (SDMA) and Non-Orthogonal Multiple Access (NOMA) in a wide range of network scenarios. Through the splitting of user messages into common and private components, RSMA has more flexible interference management with partial decoding and interference treating, while enjoying better resilience to channel uncertainties and user deployments [1]-[3].

Although there have been comprehensive investigations considering RSMA under perfect propagation conditions, its performance in rank-deficient fading channels-more specifically, keyhole channels-has not received significant consideration. Such channels represent a particular form of spatial correlation in which the MIMO channel collapses to a rank-one variant of itself, significantly diminishing spatial multiplexing gains [4]-[6]. In realistic settings, like urban environments or tunnel passages, keyhole effects occur naturally as a result of physical limitations along signal-propagation paths.

In this paper, we investigate RSMA outage performance under keyhole fading. We consider a two-user downlink system and obtain exact and asymptotic results on the outage probability for pSIC and ipSIC. We also address the achievable diversity order and illustrate the effect of the keyhole on the reliability of the system. Our findings demonstrate that RSMA achieves a significant performance improvement over NOMA, especially when the SIC is imperfect or in the medium SNR range.

1.1 Related Work

RSMA has gained increasing attention for its potential to integrate and outperform classical multiple-access schemes in various performance aspects. Earlier works have investigated RSMA in the scenario of perfect MIMO channels, where ergodic capacity, energy efficiency and user fairness have been tackled [7]-[9]. More recently, analytic works have started investigating the outage behavior of RSMA, particularly when the blocklength is finite and there are short-packet constraints [10]-[13].

Keyhole fading, on the other hand, has been a well-known severe limitation to MIMO systems, which

1. H.-N. Nguyen is with Faculty of Technology and Engineering, Saigon Uni. (SGU), Ho Chi Minh City, Vietnam. Email: nhu.nh@sgu.edu.vn
2. P.-C. Ngo is with Faculty of Electrical and Electronics Eng., Ly Tu Trong College of Ho Chi Minh (LTTC), Chi Minh City, Vietnam. Email: ngophongcuong@lttc.edu.vn

has the tendency to make the spatial-diversity gain inoperative. Theoretical keyhole channel models are usually in the form of a product of two independent fading distributions and have been investigated under Nakagami- m, Rayleigh and Ricean conditions [14]-[17]. Recent advances have also significantly highlighted the necessity of analyzing multiple access and secure transmissions specifically over generalized Nakagami- m fading environments [18]-[19]. A number of works have obtained closed-form outage or bit error rate results in such channels, but primarily for single-user or conventional SISO setups.

To date, there have been limited attempts at marrying RSMA with keyhole channel modeling. Some initial attempts have looked at RSMA over correlated or rank-deficient channels, but typically resort to simulation-based analysis or approximations [20]-[21]. In this paper, we bridge the gap by presenting a systematic analytical treatment of RSMA over keyhole channels with exact expressions, asymptotic outage behavior and diversity order under both pSIC and ipSIC assumptions.

1.2 Motivation and Contributions

While RSMA continues to evolve as a powerful transmission technique, its performance under non-ideal propagation conditions is still a largely uncharted territory. Specifically, keyhole fading-where rich scattering does not provide spatial rank gain is a particular challenge for sophisticated multi-user communication strategies. Unlike conventional SISO fading, the keyhole channel model handles the extreme spatial correlation, which reduces the rank of the channel matrix to one, regardless of the number of antennas or spatial paths. Such a phenomenon can significantly reduce the spatial multiplexing gain that RSMA normally leverages to combat interference. Motivated by this limitation, we aim to theoretically characterize the behavior of RSMA under keyhole channels. Although a number of papers have dealt with outage performance of RSMA or keyhole channels separately, their intersection has never been exactly investigated. Also missing is the role of imperfect successive interference cancellation (ipSIC) under rank-deficient conditions. Not only of theoretical interest, but also of practical importance for system implementation in tunnels, indoor corridors or device-to-device environments with unfavorable scattering, it is to investigate the performance of RSMA under this adverse fading condition. The main contributions of this paper are as follows:

- Keyhole-aware outage analysis for RSMA: We present a new analytical method to analyze the outage probability of a two-user RSMA system in keyhole fading channels represented by the product of independent Nakagami-m distributions.
- Closed-form and asymptotic solutions: We obtain exact closed-form solutions for outage probability, followed by convenient asymptotic approximations at high SNR from which we can compute the diversity order analytically.
- Imperfect SIC included: Both perfect and imperfect SIC situations are taken into account within the analysis, so that realistic evaluation of the system performance under residual interference can be done.
- Comparison with NOMA: In order to provide a performance benchmark for RSMA, we also derive and simulate the outage performance of traditional NOMA in the same environment, showing the resilience of RSMA to keyhole-affected channels. All analytical results are confirmed by extensive Monte-Carlo simulations, proving the accuracy of the derived expressions and further illustrating the performance gain of RSMA over NOMA.

1.3 Organization

The rest of this paper is structured as follows: Section 2 presents the system model, such as the keyhole channel and RSMA scheme; Section 3 calculates the outage probability for pSIC and ipSIC cases; Section 4 investigates the diversity order; Section 5 provides numerical and simulation results comparing RSMA and NOMA; and Section 6 concludes the paper.

2. SYSTEM MODEL

2.1 System Architecture

We consider a downlink multi-user single-antenna RSMA system in which a base station (BS) communicates with N legitimate single-antenna users through a keyhole of scattering cross-section δ

that separates the BS and the users as shown in Fig. 1, where the BS-to-keyhole link g_0 and the keyhole-to-user n link g_n are each modeled as independent Nakagami- m fading processes, so that the overall channel gain to user n is the product $g_0 g_n$.

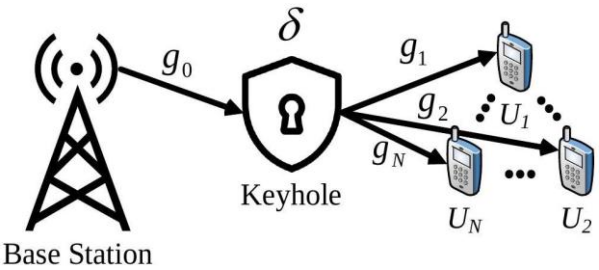


Figure 1. The system model of keyhole-based RSMA network.

2.2 Received Signals

To efficiently serve multiple users, the BS implements RSMA, which enables the simultaneous transmission of common and private messages. This strategy improves spectral efficiency and reduces inter-user interference. The transmitted signal is given by:

$$\bar{x} = \sqrt{P_S}\left(\sqrt{a_c}x_c + \sum_{n=1}^{N}\sqrt{a_n}x_n\right), \tag{1}$$

where P_S denotes the total transmitted power of the BS, x_c is the common message intended for all users with a power allocation coefficient a_c and x_n is the private message for the n^{th} user, assigned with power coefficient a_n, ensuring $a_c + \sum_{n=1}^{N} a_n = 1$.

The received signal at the n^{th} user, under the keyhole channel condition is given by:

$$\bar{y}_n = g_0\delta g_n\bar{x} + \omega_n = \underbrace{g_0\delta g_n\sqrt{a_c P_S}x_c}_{\text{Common message}} + \underbrace{g_0\delta g_n\sqrt{a_n P_S}x_n}_{\text{Desired private message}} + \underbrace{\sum_{j=1,j\neq n}^{N} g_0\delta g_n\sqrt{a_j P_S}x_j}_{\text{Interference}} + \underbrace{\omega_n}_{\text{AWGN}}, \tag{2}$$

where ω_n is the additive white Gaussian noise (AWGN) with variance N_0. To decode the intended message, each user employs SIC in a two-step decoding process:

2.2.1 Decoding the Common Message

User n first decodes the common message x_c, considering all private messages as interference. The corresponding Signal-to-Interference-plus-Noise Ratio (SINR) for decoding x_c at user n is given by:

$$\bar{\gamma}_{c,n} = \frac{a_c P_S\delta^2|g_0|^2|g_n|^2}{(1-a_c)P_S\delta^2|g_0|^2|g_n|^2 + N_0} = \frac{a_c\rho_S\delta^2|g_0|^2|g_n|^2}{(1-a_c)\rho_S\delta^2|g_0|^2|g_n|^2 + 1}. \tag{3}$$

where $\rho_S = P_S/N_0$ is the transmit signal-to-noise ratio (SNR). Note that x_c and x_n are supposed to be normalized unity power signals, i.e., $\mathbb{E}\{|x_c|^2\} = \mathbb{E}\{|x_n|^2\} = 1$ in which $\mathbb{E}\{.\}$ denotes expectation operation.

2.2.2 Decoding the Private Message

Once the common message x_c is successfully decoded and cancelled, each user proceeds to decode its respective private message x_n. The SINR expressions for the perfect and imperfect SIC (pSIC and ipSIC) cases are given by:

$$\bar{\gamma}_{p,n}^{\text{pSIC}} = \frac{a_n\rho_S\delta^2|g_0|^2|g_n|^2}{\rho_S\delta^2|g_0|^2|g_n|^2\sum_{j=1,j\neq n}^{N} a_j + 1} \tag{4a}$$

$$\bar{\gamma}_{p,n}^{\text{ipSIC}} = \frac{a_n\rho_S\delta^2|g_0|^2|g_n|^2}{\rho_S\delta^2|g_0|^2|g_n|^2\sum_{j=1,j\neq n}^{N} a_j + \varpi\rho_S|g_I|^2 + 1} \tag{4b}$$

Here, $\varpi = 0$ and $\varpi = 1$ correspond to the cases of pSIC and ipSIC, respectively. The residual interference caused by ipSIC is modeled using Rayleigh fading, where the corresponding complex

channel gain is represented as $g_I \sim \mathcal{CN}(0, \Omega_I)$, with $0 \leq \Omega_I < 1$ as suggested in [22]. In this context, $\mathcal{CN}(.,.)$denotes the complex Gaussian distribution.

2.3 Channel Distributions

When the wireless channel $g_b, b \in \{0, n\}$ follows a Nakagami- m distribution, the corresponding channel power gain $|g_b|^2$ follows a Gamma distribution. The probability density function (PDF) and cumulative distribution function (CDF) of this Gamma-distributed gain are provided below, under the assumption that the fading parameter m_b is an integer greater than or equal to one [23]-[24].

$$f_{|g_b|^2}(x) = \frac{x^{m_b-1}}{\Gamma(m_b)\Omega_b^{m_b}} e^{-\frac{x}{\Omega_b}} \tag{5}$$

$$F_{|g_b|^2}(x) = 1 - \frac{1}{\Gamma(m_b)}\Gamma\left(m_b, \Omega_b^{-1}x\right) = 1 - e^{-\frac{x}{\Omega_b}} \sum_{p=0}^{m_b-1} \frac{x^p}{p!\,\Omega_b^p}, \tag{6}$$

where $\Omega_b \triangleq \lambda_b/m_b$ in which λ_b and m_b representing the mean and integer fading factor, respectively.

$\Gamma(.,.) and \Gamma(.)$ stands for the upper incomplete Gamma function and the Gamma function, respectively. It is worth noting that the second line of Eq. (6) holds only when m_b is an integer.

Remark 1: It is worth mentioning that although the Nakagami fading parameter m can generally assume any real value $m \geq 0.5$, we constrain m_b to be an integer in our derivations. This mathematical assumption is widely adopted in the literature to express the incomplete Gamma function as a finite series, which is a pivotal step to obtain closed-form mathematical expressions. The general performance trends and insights drawn from this integer assumption remain fully applicable to scenarios with non-integer fading parameters.

On the other hand, the PDF and CDF of the product of two squared Nakagami-m random variables, $\mathcal{G}_b \triangleq |g_0|^2|g_n|^2$, is given as follows [25]:

$$f_{\mathcal{G}_b}(x) = \frac{2x^{\frac{m_0+m_n}{2}} - 1}{\Gamma(m_0)\Gamma(m_n)(\Omega_0\Omega_n)^{\frac{m_0+m_n}{2}}} K_{m_0-m_n}\left(\sqrt{\frac{4x}{\Omega_0\Omega_n}}\right), \tag{7}$$

$$F_{\mathcal{G}_b}(x) = 1 - \sum_{p=0}^{m_0-1} \frac{2}{p!\,\Gamma(m_n)}\left(\frac{x}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}} K_{m_n-p}\left(\sqrt{\frac{4x}{\Omega_0\Omega_n}}\right), \tag{8}$$

Here, $K_q(.)$ is the modified Bessel function of the second kind [[26], Eq. (8.407)] and $m_0 \in \mathbb{N}$ denotes a positive integer fading parameter.

Additionally, Rayleigh-distributed RVs of $|g_I|^2$ have exponential distributions with $f_{|g_I|^2}(x) = \frac{1}{\Omega_I} e^{-\frac{x}{\Omega_I}}$ and $F_{|g_I|^2}(x) = 1 - e^{-\frac{x}{\Omega_I}}$ in [27].

3. OUTAGE PROBABILITY

In RSMA-based transmission, each user receives a superposition of the common message, its own private message and the private messages intended for other users. The decoding process follows a two-step approach as described in (3), (4a) and (4b). If the SINRs for decoding the common and private messages fall below the respective thresholds $\gamma_{th}^{c,n}$ and $\gamma_{th}^{p,n}$, the link between the BS and the n^{th} user, affected by the keyhole channel condition, is considered to be in outage. In this context, $\gamma_{th}^{c,n} = 2^{R_{c,n}} - 1$ and $\gamma_{th}^{p,n} = 2^{R_{p,n}} - 1$ denote the corresponding SINR thresholds, while $R_{c,n}$ and $R_{p,n}$ represent the target rates for decoding the common and private messages, respectively.

3.1 Outage Probability with pSIC

Theorem 1. The outage probability of the BS-to- n^{th} user link under keyhole fading with pSIC is given by:

$$\mathcal{P}_{U_n}^{\text{pSIC}} = \begin{cases} 1 - \sum_{p=0}^{m_0-1} \frac{2}{p!\,\Gamma(m_n)} \left(\frac{\tilde{\gamma}_{th}^{p,n}}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}} K_{m_n-p}\left(\sqrt{\frac{4\tilde{\gamma}_{th}^{p,n}}{\Omega_0\Omega_n}}\right), & \text{if } \tilde{\gamma}_{th}^{c,n} \le \tilde{\gamma}_{th}^{p,n}, \\ 1 - \sum_{p=0}^{m_0-1} \frac{2}{p!\,\Gamma(m_n)} \left(\frac{\tilde{\gamma}_{th}^{c,n}}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}} K_{m_n-p}\left(\sqrt{\frac{4\tilde{\gamma}_{th}^{c,n}}{\Omega_0\Omega_n}}\right), & \text{if } \tilde{\gamma}_{th}^{c,n} > \tilde{\gamma}_{th}^{p,n}, \end{cases} \tag{9}$$

where $\tilde{\gamma}_{th}^{c,n} = \gamma_{th}^{c,n}/\delta^2\rho_S[a_c - (1-a_c)\gamma_{th}^{c,n}]$ and $\tilde{\gamma}_{th}^{p,n} = \gamma_{th}^{p,n}/\delta^2\rho_S[a_n - (1-a_c-a_n)\gamma_{th}^{p,n}]$. Moreover, the result in (9) is derived under the conditions $a_c > \gamma_{th}^{c,n}/(1+\gamma_{th}^{c,n})$ and $a_n > (1-a_c)\gamma_{th}^{p,n}/(1+\gamma_{th}^{p,n})$.

Proof. The outage probability at the n^{th} user with pSIC is given by:

$$\mathcal{P}_{U_n}^{\text{pSIC}} = \Pr(\bar{\gamma}_{c,n} < \gamma_{th}^{c,n} \cup \bar{\gamma}_{p,n}^{\text{pSIC}} < \gamma_{th}^{p,n}) = 1 - \Pr(\bar{\gamma}_{c,n} > \gamma_{th}^{c,n}, \bar{\gamma}_{p,n}^{\text{pSIC}} > \gamma_{th}^{p,n}). \tag{10}$$

Plugging $\bar{\gamma}_{c,n}$ and $\bar{\gamma}_{p,n}^{\text{pSIC}}$ from (3) and (4a) in (6), we have:

$$\mathcal{P}_{U_n}^{\text{pSIC}} = 1 - \Pr\left(\frac{a_c\rho_S\delta^2\mathcal{G}_b}{(1-a_c)\rho_S\delta^2\mathcal{G}_b + 1} > \gamma_{th}^{c,n}, \frac{a_n\rho_S\delta^2\mathcal{G}_b}{\rho_S\delta^2\mathcal{G}_b\sum_{j=1,j\neq n}^{N} a_j + 1} > \gamma_{th}^{p,n}\right). \tag{11}$$

After performing algebraic manipulations, Equation (11) can be rewritten as:

$$\mathcal{P}_{U_n}^{\text{pSIC}} = 1 - \Pr(\mathcal{G}_b > \tilde{\gamma}_{th}^{c,n}, \mathcal{G}_b > \tilde{\gamma}_{th}^{p,n}) = 1 - \Pr(\mathcal{G}_b > \hat{\gamma}_{th}) = F_{\mathcal{G}_b}(\hat{\gamma}_{th}), \tag{12}$$

where $\hat{\gamma}_{th} = \max(\tilde{\gamma}_{th}^{c,n}, \tilde{\gamma}_{th}^{p,n})$. Combining (8) into (12), (9) can be obtained and the proof is completed.

3.2 Outage Probability with ipSIC

From (3) and (4b), the outage probability with ipSIC can be calculated by:

$$\begin{aligned} \mathcal{P}_{U_n}^{\text{ipSIC}} &= \Pr(\bar{\gamma}_{c,n} < \gamma_{th}^{c,n} \cup \bar{\gamma}_{p,n}^{\text{pSIC}} < \gamma_{th}^{p,n}) = 1 - \Pr(\bar{\gamma}_{c,n} > \gamma_{th}^{c,n}, \bar{\gamma}_{p,n}^{\text{pSIC}} > \gamma_{th}^{p,n}) \\ &= 1 - \Pr\left(\frac{\frac{a_c\rho_S\delta^2\mathcal{G}_b}{(1-a_c)\rho_S\delta^2\mathcal{G}_b + 1} > \gamma_{th}^{c,n}}{\rho_S\delta^2\mathcal{G}_b\sum_{j=1,j\neq n}^{N} a_j + \varpi\rho_S|g_I|^2 + 1} > \gamma_{th}^{p,n}\right) \end{aligned} \tag{13}$$

Theorem 2. The outage probability of the BS-to- n^{th} user link under keyhole fading with ipSIC is expressed as:

$$\mathcal{P}_{U_n}^{\text{ipSIC}} \approx \left[1 - \sum_{p=0}^{m_0-1} \frac{2}{p!\Gamma(m_n)} \left(\frac{T_c}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}} K_{m_n-p}\left(\sqrt{\frac{4T_c}{\Omega_0\Omega_n}}\right)\right]\left(1 - e^{-\frac{x_0}{\Omega_I}}\right) + e^{\frac{1}{\varpi\rho\Omega_I}}\left[e^{-\varphi_I T_c} - \right.$$

$$\left. \sum_{p=0}^{m_0-1}\sum_{q=1}^{Q} \frac{2\pi^2\varphi_I\sqrt{1-\xi_q^2}}{p!4Q\Gamma(m_n)} \left(\frac{1}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}} \sec^2\left(\frac{\pi(\xi_q+1)}{4}\right) \times \Delta(\xi_q)^{\frac{p+m_n}{2}} e^{-\varphi_I\Delta(\xi_q)} K_{m_n-p}\left(\sqrt{\frac{4\Delta(\xi_q)}{\Omega_0\Omega_n}}\right)\right] \tag{14}$$

where $T_c = \frac{\gamma_{c,th}}{\rho_S\delta^2[a_c-(1-a_c)\gamma_{c,th}]}$, $\varphi_I = \frac{1}{\varpi\rho_S\tilde{\gamma}_{th}^{p,n}\Omega_I}$, $\Delta(x) = \tan\left(\frac{\pi(x+1)}{4}\right) + T_c$, $\xi_q = \cos\left(\frac{2q-1}{2Q}\pi\right)$ and Q is a complexity-accuracy tradeoff parameter. In this paper, we verified through MATLAB simulations that setting $Q = 200$ is sufficient to achieve excellent convergence, providing a tight match between the exact integration and the analytical approximation.

Proof. The proof is provided in Appendix A.

3.3 Outage Probability for Baseline NOMA

For comparison, a conventional two-user NOMA scheme is considered where the base station superimposes the private messages directly without a common message ($a_c = 0$). Assuming User 1 and User 2 are the near and far users, respectively, the power allocation coefficients satisfy $a_1 < a_2$ with $a_1 + a_2 = 1$. The received SINR for User 2 to decode its own message is given by:

$$\gamma_2^{\text{NOMA}} = \frac{a_2\rho_S\delta^2\mathcal{G}_2}{a_1\rho_S\delta^2\mathcal{G}_2 + 1}. \tag{15}$$

User 1 performs SIC to decode User 2's message before decoding its own. The SINR for User 1 to decode User 2's message and its own message are, respectively, formulated as:

$$\gamma_{1\to2}^{\text{NOMA}} = \frac{a_2\rho_S\delta^2\mathcal{G}_1}{a_1\rho_S\delta^2\mathcal{G}_1 + 1}, \gamma_1^{\text{NOMA}} = a_1\rho_S\delta^2\mathcal{G}_1. \tag{16}$$

Consequently, the exact outage probabilities for User 1 and User 2 under the keyhole environment can be explicitly derived as:

$$\mathcal{P}_{U_1}^{\text{NOMA}} = 1 - \Pr\left(\gamma_{1\to2}^{\text{NOMA}} \geq \gamma_{th}^2, \gamma_1^{\text{NOMA}} \geq \gamma_{th}^1\right) = F_{\mathcal{G}_1}\left(\max(\tau_1, \tau_2)\right), \tag{17}$$

$$\mathcal{P}_{U_2}^{\text{NOMA}} = \Pr\left(\gamma_2^{\text{NOMA}} < \gamma_{th}^2\right) = F_{\mathcal{G}_2}(\tau_2), \tag{18}$$

where $\tau_2 = \frac{\gamma_{th}^2}{\rho_S\delta^2(a_2 - a_1\gamma_{th}^2)}$ under the condition $a_2 > a_1\gamma_{th}^2$ (otherwise $\mathcal{P}_{U_2}^{\text{NOMA}} = 1$), $\tau_1 = \frac{\gamma_{th}^1}{\rho_S\delta^2 a_1}, \gamma_{th}^n = 2^{R_{p,n}} - 1$ is the target threshold for User n and $F_{\mathcal{G}_b}(.)$ is the CDF of the product of two Nakagami-m random variables given in (8).

4. DIVERSITY ANALYSIS

To gain further insights, the diversity order reached by users in two situations can be determined in this section using the analytical data presented above. The diversity order is defined as [28], Eq. (18).

$$d_{\mathrm{U}_n}^{\star} = -\lim_{\rho_S\to\infty} \frac{\log\left(\mathcal{P}_{U_n}^{\infty,\star}(\rho_S)\right)}{\log\rho_S}, \star \in \{\text{pSIC}, \text{ipSIC}\}. \tag{19}$$

With the help of [29], the asymptotic expression for the outage probability at U_n under pSIC in the high-SNR regime ($\rho_S \to \infty$) can be formulated as:

$$\mathcal{P}_{\mathrm{U}_n}^{\infty,\text{pSIC}} \underset{\rho_S\to\infty}{\approx} \frac{\psi_n(4\hat{\gamma}_{th})^{m_0}}{2m_0(\Omega_0\Omega_n)^{m_0}}, \tag{20}$$

where the condition $|m_0 - m_n| = 0.5$ is mathematically adopted from [29] to facilitate the exact simplification of the associated mathematical functions (e.g., modified Bessel function) into an incomplete Gamma function. This specific half-integer difference is required to explicitly extract a tight asymptotic bound and determine the diversity order in the high-SNR regime and $\psi_n = \frac{2^{1-2m_0}\sqrt{\pi}}{\Gamma(m_0)\Gamma(m_n)}$. By substituting (20) into (19) and performing algebraic simplifications, the diversity order of U_n with pSIC is found to be equal to m_0. For ipSIC, when ρ_S goes to infinity, then we have $\bar{\gamma}_{c,n} \approx \frac{a_c}{1-a_c}$ and $\bar{\gamma}_{p,n}^{\text{ipSIC}} \approx \frac{a_n\delta^2|g_0|^2|g_n|^2}{\delta^2|g_0|^2|g_n|^2\sum_{j=1,j\neq n}^{N} a_j + \varpi|g_I|^2}$, the asymptotic expression for $\mathcal{P}_{\mathrm{U}_n}^{\infty,\text{ipSIC}}$ is calculated as:

$$\mathcal{P}_{\mathrm{U}_n}^{\infty,\text{ipSIC}} = \begin{cases} 1, & if\ \dfrac{a_c}{1-a_c} < \gamma_{th}^{c,n} \\ \Pr\left(\dfrac{a_n\delta^2|g_0|^2|g_n|^2}{\delta^2|g_0|^2|g_n|^2\sum_{j=1,j\neq n}^{N} a_j + \varpi|g_I|^2} < \gamma_{th}^{p,n}\right), & \text{otherwise} \end{cases} \tag{21}$$

It is noted that we can rewrite (21) as:

$$\mathcal{P}_{\mathrm{U}_n}^{\infty,\text{ipSIC}} = \Pr(\mathcal{G}_b < \phi_n|g_I|^2) = \frac{1}{\Omega_I}\int_0^{\infty} F_{\mathcal{G}_b}(\phi_n x)e^{-\frac{x}{\Omega_I}}dx \tag{22}$$

where $\phi_n = \frac{\gamma_{th}^{p,n}\varpi}{\delta^2\left(a_n - \gamma_{th}^{p,n}\sum_{j=1,j\neq n}^{N} a_j\right)}$. Now, invoking (8) into (22), we have $\mathcal{P}_{\mathrm{U}_n}^{\infty,\text{ipSIC}}$ is given by:

$$\begin{aligned}\mathcal{P}_{\mathrm{U}_n}^{\infty,\text{ipSIC}} &= \frac{1}{\Omega_I}\int_0^{\infty}\left[1 - \sum_{p=0}^{m_0-1}\frac{2}{p!\,\Gamma(m_n)}\left(\frac{\phi_n x}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}} K_{m_n-p}\left(\sqrt{\frac{4\phi_n x}{\Omega_0\Omega_n}}\right)\right]e^{-\frac{x}{\Omega_I}}dx \\ &= 1 - \sum_{p=0}^{m_0-1}\frac{2\phi_n^{\frac{p+m_n}{2}}}{p!\,\Gamma(m_n)\Omega_I(\Omega_0\Omega_n)^{\frac{p+m_n}{2}}}\int_0^{\infty} x^{\frac{p+m_n}{2}}e^{-\frac{x}{\Omega_I}}K_{m_n-p}\left(\sqrt{\frac{4\phi_n x}{\Omega_0\Omega_n}}\right)dx\end{aligned} \tag{23}$$

Exchanging the variable $t = \frac{4}{\pi}\arctan(x) - 1 \Rightarrow \tan\left(\frac{\pi(t+1)}{4}\right) = x \Rightarrow \frac{\pi}{4}\sec^2\left(\frac{\pi}{4}(t+1)\right)dt = dx$, we have $\mathcal{C}_1^{ipSIC}$ is given by:

$$\mathcal{P}_{\mathrm{U}_n}^{\infty,\mathrm{ipSIC}} = 1 - \sum_{p=0}^{m_0-1} \frac{\pi \phi_n^{\frac{p+m_n}{2}}}{p!\, 2\Gamma(m_n)\Omega_I(\Omega_0\Omega_n)^{\frac{p+m_n}{2}}} \int_0^{\infty} \sec\left(\frac{\pi(t+1)}{4}\right) \Theta(t)^{\frac{p+m_n}{2}} e^{-\frac{\Theta(t)}{\Omega_I}} \times K_{m_n-p}\left(\sqrt{\frac{4\phi_n\Theta(t)}{\Omega_0\Omega_n}}\right) dt$$

where $\Theta(t) = \tan\left(\frac{\pi(t+1)}{4}\right)$.

By applying the Gaussian-Chebyshev quadrature, (24) can be approximated as:

$$\mathcal{P}_{\mathrm{U}_n}^{\infty,\mathrm{ipSIC}} \approx 1 - \sum_{p=0}^{m_0-1} \frac{\pi^2 \phi_n^{\frac{p+m_n}{2}}}{p!2Q\Gamma(m_n)\Omega_I(\Omega_0\Omega_n)^{\frac{p+m_n}{2}}} \sum_{q=1}^{Q} \sqrt{1-\xi_q^2} \sec\left(\frac{\pi(\xi_q+1)}{4}\right) \times \Theta(\xi_q)^{\frac{p+m_n}{2}} e^{-\frac{\Theta(\xi_q)}{\Omega_I}} K_{m_n-p}\left(\sqrt{\frac{4\phi_n\Theta(\xi_q)}{\Omega_0\Omega_n}}\right) \quad (24)$$

By incorporating (24) into (19) and following a series of simplifications, it is possible to derive the diversity order of U_n in the context of ipSIC, which is determined to be 0, a finding that is further corroborated by the graphical representations presented in Section 6.

5. NUMERICAL RESULTS

In this Section, we present the simulated outcomes pertaining to the analyzed system and juxtapose them with analytical formulations to validate their accuracy. Absent any specific indications, we adopt the system parameters delineated below: $a_c = 0.4, R_{c,n} = 0.2, R_{p,1} = 0.2, R_{p,2} = 0.1$ and $\delta = 0.5$. Additionally, we examine a straightforward scenario with $N = 2$ with power distribution coefficients explicitly set as $a_1 = 0.24$ and $a_2 = 0.36$, ensuring the total power constraint $a_c + a_1 + a_2 = 1$ is strictly satisfied. For the channel gains, the mean values are set to $\lambda_0 = \lambda_1 = \lambda_2 = 1$. The selection of unity mean values corresponds to a normalized fading channel standard, which is widely adopted in the existing literature (e.g., [25],[28]) to evaluate baseline theoretical limits. This symmetric channel setup ensures that the performance comparison between RSMA and NOMA is evaluated under a generalized fading environment. The residual interference variance is set to $\Omega_I = 0.01$. Furthermore, the Gauss-Chebyshev parameter is chosen as $Q = 200$ to achieve a precise approximation [30].

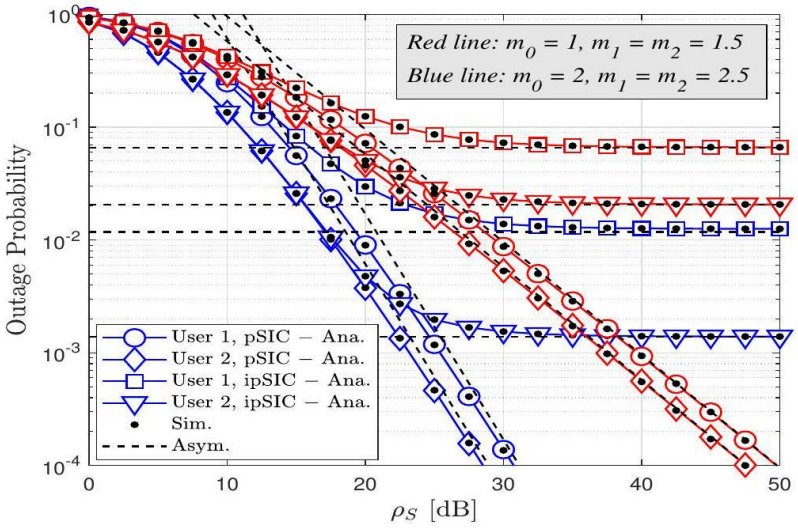


Figure 2. Outage performance *versus* ρ_S for $\mathrm{U_n}$.

Figure 2 shows the outage probability for two users under two sets of Nakagami-m parameters ($m_0 = 1, m_1 = m_2 = 1.5$ *vs.* $m_0 = 2, m_1 = m_2 = 2.5$) shows that increasing the fading severity parameters

consistently lowers the outage across all SNR values. At low SNR (10-20 dB), both curves drop sharply, indicating that even moderate increases in transmit power yield significant reliability gains despite the keyhole effect. In the high-SNR region (above 30 dB), both curves begin to flatten, reflecting the limited diversity imposed by the rank-one keyhole channel higher m values delay, but do not eliminate this outage floor.

Figure 3 depicts the comparison of RSMA and NOMA, RSMA maintains a clear outage advantage for both User 1 and User 2 across the entire SNR range, under both perfect and imperfect SIC assumptions. The performance gap widens in the mid-to-high SNR regime ($25-40$ dB), illustrating RSMA's superior interference mitigation under keyhole fading even a small amount of residual ipSIC interference hurts NOMA more severely.

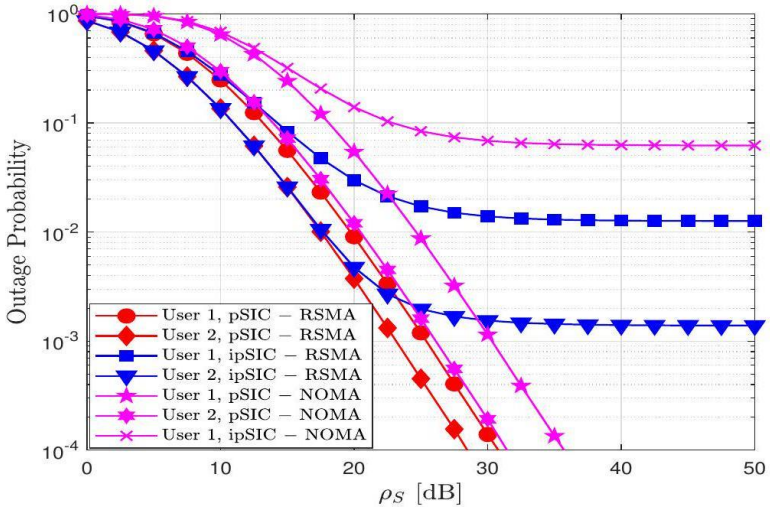


Figure 3. Comparison between RSMA and NOMA for the outage probability *versus* ρ_S.

Figure 4 shows outage probability against the common-message power coefficient a_c at two SNR values (20 dB dashed, 30 dB solid). It reveals an optimal allocation window around $a_c \approx 0.4-0.6$. Outside this window, the outage rises markedly: too little a_c starves the common stream and too much a_c reduces private-stream SINR. Increasing SNR shifts both curves downward, but preserves the same optimal region, confirming that power-split tuning remains critical under keyhole conditions.

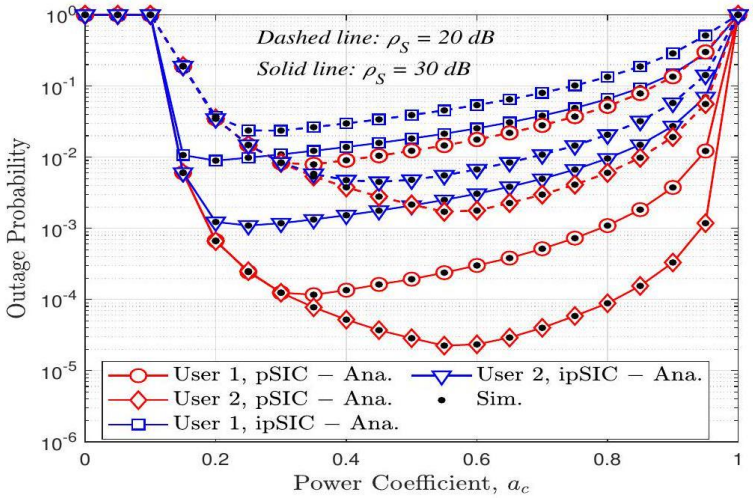


Figure 4. The outage probability versus power coefficient a_c with $m_0 = 2$ and $m_1 = m_2 = 2.5$.

Figure 5 depicts the outage probability *versus* the keyhole parameter δ at 30 dB. It displays a steep decline as δ increases (i.e., the keyhole attenuates less), demonstrating how stronger keyhole severity ($\delta < 0.2$) forces outage rates above 10^{-2}. For $\delta > 0.6$, the outage for higher Nakagami- m values ($m_0 = 3, m_1 = m_2 = 3.5$) plunges below 10^{-4}, indicating that RSMA benefits substantially from weaker keyhole effects and higher fading parameters. The gap between the two Nakagami- m curves also underscores how fading severity and keyhole strength jointly govern reliability.

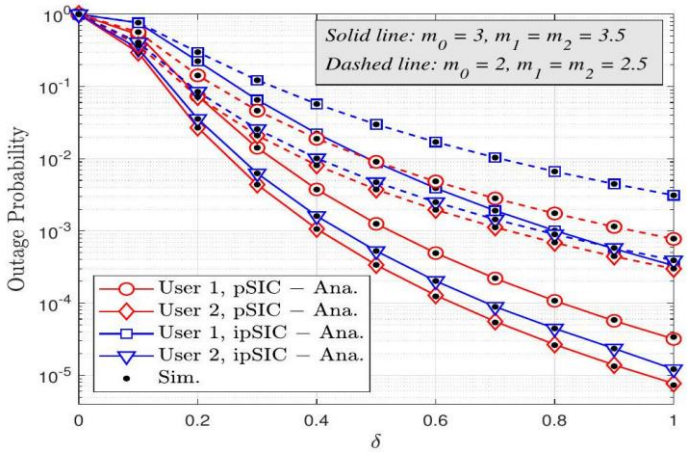


Figure 5. The outage probability *versus* the impact of the keyhole parameter δ with $\rho_S = 30$ dB.

6. CONCLUSIONS

This paper provides a unified analytical framework to assess the performance of RSMA in keyhole fading channels. Exact and asymptotic outage probabilities are established under perfect and imperfect SIC assumptions. The findings show that RSMA is more robust against the keyhole effect than traditional NOMA, especially in the presence of residual interference. Our analysis of diversity also shows that the keyhole channel degradation restricts the diversity attainable, but RSMA is resilient because of its adaptable message splitting framework. The expressions thus obtained agree well with simulation results and the theoretical analysis is therefore validated. The results offer a better understanding of RSMA performance under rank-deficient channels caused by spatial correlation or keyhole effects and support its applicability to practical wireless systems subject to spatial correlation.

Appendix A: Proof of Theorem 2

From (13), the outage probability occurs if either decoding step fails, i.e.

$$\mathcal{P}_{U_n}^{\text{ipSIC}} = 1 - \Pr\left(\mathcal{G}_b > \tilde{\gamma}_{th}^{c,n}, \mathcal{G}_b > \tilde{\gamma}_{th}^{p,n}(\varpi\rho_S|g_I|^2 + 1)\right) = 1 - \Pr\left(\mathcal{G}_b > \max\left(T_c, T_p(|g_I|^2)\right)\right), \quad (A.1)$$

Define thresholds

$$T_c \triangleq \tilde{\gamma}_{th}^{c,n} = \frac{\gamma_{c,th}}{\rho_S\delta^2\left[a_c - (1-a_c)\gamma_{c,th}\right]} \quad (A.2)$$

$$T_p(x) \triangleq \tilde{\gamma}_{th}^{p,n}(1 + \varpi\rho_S x) = \frac{\gamma_{p,th}}{\rho_S\delta^2\left[a_n - \gamma_{p,th}\sum_{j\neq n} a_j\right]}(1 + \varpi\rho_S x) \quad (A.3)$$

We have:

$$\max\left(T_c, T_p(x)\right) = \begin{cases} T_c, & T_p(x) \le T_c, \\ T_p(x), & T_p(x) > T_c. \end{cases} \quad (A.4)$$

Let $x = |g_I|^2$. By solving $T_p(x) = T_c$, we obtain:

$$T_p(x) \le T_c \Rightarrow x \le T_0 = \frac{T_c/T_p(0) - 1}{\varpi\rho_S} = \frac{\tilde{\gamma}_{th}^{c,n} - \tilde{\gamma}_{th}^{p,n}}{\varpi\rho_S\tilde{\gamma}_{th}^{p,n}}. \quad (A.5)$$

Then, outage probability becomes:

$$\mathcal{P}_{U_n}^{\text{ipSIC}} = \int_0^\infty F_{\mathcal{G}_b}\left(\max\left(T_c, T_p(x)\right)\right) f_{|g_I|^2}(x)dx = \underbrace{\frac{1}{\Omega_I}\int_0^{T_0} F_{\mathcal{G}_b}(T_c)e^{-\frac{x}{\Omega_I}}dx}_{I_1} + \underbrace{\frac{1}{\Omega_I}\int_{T_0}^\infty F_{\mathcal{G}_b}\left(T_p(x)\right)e^{-\frac{x}{\Omega_I}}dx}_{I_2} \quad (A.6)$$

where $f_x(x) = e^{-x/\Omega_I}/\Omega_I$ and $F_{\mathcal{G}_b}(\cdot)$ is the CDF of two squared Nakagami- m RVs. From (8), we have I_1 is calculated as:

$$I_1 = F_{\mathcal{G}_b}(T_c)\left(1 - e^{-\frac{x_0}{\Omega_I}}\right) = \left[1 - \sum_{p=0}^{m_0-1}\frac{2}{p!\,\Gamma(m_n)}\left(\frac{T_c}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}} K_{m_n-p}\left(\sqrt{\frac{4T_c}{\Omega_0\Omega_n}}\right)\right]\left(1 - e^{-\frac{x_0}{\Omega_I}}\right). \quad (A.7)$$

For I_2, use the change of variables $y = T_p(x)$ to obtain:

$$I_2 = \varphi_I e^{\frac{1}{\varpi\rho_S\Omega_I}}\left[\frac{e^{-\varphi_I T_c}}{\varphi_I} - \sum_{p=0}^{m_0-1}\frac{2}{p!\,\Gamma(m_n)}\left(\frac{1}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}} I_3\right], \quad (A.8)$$

where $\varphi_I = \frac{1}{\varpi\rho_S\tilde{\gamma}_{th}^{p,n}\Omega_I}$ and $I_3 = \int_{T_c}^\infty y^{\frac{p+m_n}{2}} K_{m_n-p}\left(\sqrt{\frac{4y}{\Omega_0\Omega_n}}\right)e^{-\varphi_I y}dy$.

Let $t = \frac{4}{\pi}\arctan(y - T_c) - 1 \Rightarrow y = \tan\left((t+1)\frac{\pi}{4}\right) + T_c \Rightarrow dy = \frac{\pi}{4}\sec^2\left((t+1)\frac{\pi}{4}\right)$, I_3 is given as:

$$I_3 = \frac{\pi}{4}\int_{-1}^{1}\sec^2\left(\frac{\pi(t+1)}{4}\right)\Delta(t)^{\frac{p+m_n}{2}} K_{m_n-p}\left(\sqrt{\frac{4\Delta(t)}{\Omega_0\Omega_n}}\right)e^{-\varphi_I\Delta(t)}dt, \quad (A.9)$$

where $\Delta(t) = \tan\left(\frac{\pi(t+1)}{4}\right) + T_c$.

Unfortunately, finding a closed-form expression for (A.9) is a tough task, but an accurate approximation can be obtained for it. By using Gaussian-Chebyshev quadrature [[31], Eq. (25.4.38)], (A.9) can be achieved.

$$I_3 \approx \frac{\pi^2}{4Q}\sum_{q=1}^{Q}\sqrt{1-\xi_q^2}\sec^2\left(\frac{\pi(\xi_q+1)}{4}\right)\Delta(\xi_q)^{\frac{p+m_n}{2}}e^{-\varphi_I\Delta(\xi_q)}K_{m_n-p}\left(\sqrt{\frac{4\Delta(\xi_q)}{\Omega_0\Omega_n}}\right), \quad (A.10)$$

where $\xi_q = \cos\left(\frac{2q-1}{2Q}\pi\right)$ and Q is a complexity-accuracy tradeoff parameter. Substituting (A.10) into (A.8), I_2 can be obtained as:

$$I_2 \approx \varphi_I e^{\frac{1}{\varpi\rho_S\Omega_I}}\left[\frac{e^{-\varphi_I T_c}}{\varphi_I} - \sum_{p=0}^{m_0-1}\sum_{q=1}^{Q}\frac{2\pi^2\sqrt{1-\xi_q^2}}{p!4Q\Gamma(m_n)}\left(\frac{1}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}}\sec^2\left(\frac{\pi(\xi_q+1)}{4}\right)\times\Delta(\xi_q)^{\frac{p+m_n}{2}}e^{-\varphi_I\Delta(\xi_q)}K_{m_n-p}\left(\sqrt{\frac{4\Delta(\xi_q)}{\Omega_0\Omega_n}}\right)\right] \quad (A.11)$$

Substituting (A.11) and (A.7) into (A.6) and applying the integration by parts, $\mathcal{P}_{U_n}^{\text{ipSIC}}$ can be re-expressed as:

$$\mathcal{P}_{U_n}^{\text{ipSIC}} \approx \left[1 - \sum_{p=0}^{m_0-1}\frac{2}{p!\,\Gamma(m_n)}\left(\frac{T_c}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}}K_{m_n-p}\left(\sqrt{\frac{4T_c}{\Omega_0\Omega_n}}\right)\right]\left(1-e^{-\frac{x_0}{\Omega_I}}\right)$$

$$+e^{\frac{1}{\varpi\rho_S\Omega_I}}\left[e^{-\varphi_I T_c} - \sum_{p=0}^{m_0-1}\sum_{q=1}^{Q}\frac{2\pi^2\varphi_I\sqrt{1-\xi_q^2}}{p!4Q\Gamma(m_n)}\left(\frac{1}{\Omega_0\Omega_n}\right)^{\frac{p+m_n}{2}}\sec^2\left(\frac{\pi(\xi_q+1)}{4}\right)\times\Delta(\xi_q)^{\frac{p+m_n}{2}}e^{-\varphi_I\Delta(\xi_q)}K_{m_n-p}\left(\sqrt{\frac{4\Delta(\xi_q)}{\Omega_0\Omega_n}}\right)\right]. \quad (A.12)$$

This completes the proof.

REFERENCES

[1] W. Jaafar et al., "On the Downlink Performance of RSMA-based UAV Communications," IEEE Trans. on Vehicular Technology, vol. 69, no. 12, pp. 16258-16263, 2020.

[2] F. Xiao et al., "Outage Performance Analysis of RSMA-aided Semi-grant-free Transmission Systems," IEEE Open J. of the Communications Society, vol. 4, pp. 253-268, 2023.

[3] T.-H. Vu et al., "On Performance of Downlink THz-based Rate-splitting Multiple-access (RSMA): Is It Always Better Than NOMA?," IEEE Trans. on Vehicular Techn., vol. 73, no. 3, pp. 4435-4440, 2024.

[4] P. Almers, f. Tufvesson and A. F. Molisch, "Keyhole Effect in MIMO Wireless Channels: Measurements and Theory," IEEE Trans. on Wireless Comm., vol. 5, no. 12, pp. 3596-3604, 2006.

[5] H. Zhang et al., "Performance Analysis of MIMO-HARQ Assisted V2V Communications with Keyhole Effect," IEEE Trans. on Comm., vol. 70, no. 5, pp. 3034-3046, May 2022.

[6] X. Zang et al., "On the Secrecy Performance of MIMOME Keyhole Channels Aided with Artificial Noise," IEICE Trans. on Comm., vol. E108-B, no. 5, pp. 631-639, May 2025.

[7] A. Krishnamoorthy and R. Schober, "Downlink MIMO-RSMA with Successive Null-space Precoding," IEEE Trans. on Wireless Communications, vol. 21, no. 11, pp. 9170-9185, 2022.

[8] O. Dizdar et al., "RSMA for Overloaded MIMO Networks: Low Complexity Design for Max-min Fairness," IEEE Trans. on Wireless Comm., vol. 23, no. 6, pp. 6156-6173, Jun. 2024.

[9] Y. Zhang et al., "Enhancing Secrecy in Hardware Impaired Cell-free Massive MIMO by RSMA," IEEE Trans. on Wireless Comm., vol. 23, no. 12, pp. 18788-18805, Dec. 2024.

[10] Y. Xu, Y. Mao, O. Dizdar and B. Clerckx, "Rate-Splitting Multiple Access With Finite Block Length for Short-packet and Low-latency Downlink Communications," IEEE Trans. on Vehicular Technology, vol. 71, no. 11, pp. 12333-12337, Nov. 2022.

[11] M. Katwe, K. Singh, B. Clerckx and C.-P. Li, "Rate Splitting Multiple Access for Energy Efficient RIS-aided Multi-user Short-packet Communications," Proc. of IEEE Global Communications Conf. (GLOBECOM) Workshops, pp. 644-649, Rio de Janeiro, Brazil, 2022.

[12] S. K. Singh, K. Agrawal, K. Singh, B. Clerckx and C.-P. Li, "RSMA Enhanced RIS-FD-UAV Aided Short Packet Communications under Imperfect SIC," Proc. of IEEE Global Communications Conf. (GLOBECOM) Workshops, pp. 1549-1554, Rio de Janeiro, Brazil, 2022.

[13] T.-H. Vu et al., "Rate-splitting Multiple Access-assisted THz-based Short-packet Communications," IEEE Wireless Communications Letters, vol. 12, no. 12, pp. 2218-2222, Dec. 2023.

[14] H. Shin and J. H. Lee, "Performance Analysis of Space-time Block Codes over Keyhole Nakagamim Fading Channels," IEEE Transactions on Vehicular Technology, vol. 53, no. 2, pp. 351-362, 2004.

[15] N. H. Tran, H. H. Nguyen and T. Le-Ngoc, "Performance Analysis and Design Criteria of BICMID with Signal Space Diversity for Keyhole Nakagami-m Fading Channels," IEEE Trans. on Information Theory, vol. 55, no. 4, pp. 1592-1602, Apr. 2009.

[16] C. Zhong et al., "Ergodic Mutual Information Analysis for Multi Keyhole MIMO Channels," IEEE Trans. on Wireless Communications, vol. 10, no. 6, pp. 1754-1763, Jun. 2011.

[17] L. K. S. Jayasinghe, N. Rajatheva, P. Dharmawansa and M. Latva-Aho, "Non-coherent Amplify-and Forward MIMO Relaying with OSTBC Over Rayleigh-Rician Fading Channels," IEEE Trans. on Vehicular Technology, vol. 62, no. 4, pp. 1610-1622, May 2013.

[18] A. M. Magableh et al., "Performance of Non-orthogonal Multiple Access (NOMA) Systems over N-Nakagami-m Multipath Fading Channels for 5G and Beyond," IEEE Trans. on Vehicular Techn., vol. 71,

no. 11, pp. 11615-11623, Nov. 2022.
[19] B. Ji et al., "Research on Secure Transmission Performance of Electric Vehicles under Nakagami-m Channel," IEEE Trans. on Intelligent Transportation Systems, vol. 22, no. 3, pp. 1881-1891, 2021.
[20] Z. Yang et al., "Optimization of Rate Allocation and Power Control for Rate Splitting Multiple Access (RSMA)," IEEE Transactions on Communications, vol. 69, no. 9, pp. 5988-6002, Sep. 2021.
[21] A. Krishnamoorthy and R. Schober, "Downlink Massive MU-MIMO with Successively-regularized Zero Forcing Precoding," IEEE Wireless Communications Letters, vol. 12, no. 1, pp. 114-118, 2023.
[22] X. Yue and Y. Liu, "Performance Analysis of Intelligent Reflecting Surface Assisted NOMA Networks," IEEE Trans. on Wireless Communications, vol. 21, no. 4, pp. 2623-2636, Apr. 2022.
[23] D.-T. Do et al., "UAV Relaying Enabled NOMA Network With Hybrid Duplexing and Multiple Antennas," IEEE Access, vol. 8, pp. 186993-187007, 2020.
[24] D.-T. Do et al., "Antenna Selection and Device Grouping for Spectrum-efficient UAV-Assisted IoT Systems," IEEE Internet of Things Journal, vol. 10, no. 9, pp. 8014-8030, May 2023.
[25] M. Elsayed et al., "Symbiotic Ambient Backscatter IoT Transmission Over NOMA-enabled Network," Proc. of IEEE Int. Conf. on Communi. (ICC), pp. 1549-1554, Seoul, S. Korea, 2022.
[26] I. S. Gradshteyn and I. M. Ryzhik, Table of Integrals, Series and Products, ISBN-10: 0-12-373637-4, Academic Press, 2014.
[27] T.-T. T. Dao, S. Q. Nguyen, H. N. Nguyen, P. X. Nguyen and Y.-H. Kim, "Performance Evaluation of Downlink Multiple Users NOMA-Enable UAV-aided Communication Systems Over Nakagami-m Fading Environments," IEEE Access, vol. 9, pp. 151641-151653, 2021.
[28] X. Yue, Y. Liu, S. Kang, A. Nallanathan and Z. Ding, "Exploiting Full/Half-duplex User Relaying in NOMA Systems," IEEE Trans. on Communications, vol. 66, no. 2, pp. 560-575, Feb. 2018.
[29] N. Jaiswal and N. Purohit, "Performance Analysis of NOMA-enabled Vehicular Communication Systems with Transmit Antenna Selection over Double Nakagami-m Fading," IEEE Trans. on Vehicular Technology, vol. 70, no. 12, pp. 12725-12741, Dec. 2021.
[30] Y. Cheng, K. H. Li, Y. Liu, K. C. Teh and H. V. Poor, "Downlink and Uplink Intelligent Reflecting Surface Aided Networks: NOMA and OMA," IEEE Trans. on Wireless Communications, vol. 20, no. 6, pp. 3988-4000, Jun. 2021.
[31] M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables, vol. 55, US Government Printing Office, 1948.

ملخص البحث:

تتناول هذه الورقة البحثية الأداء النّظري لتقنية الوصول المتعدّد بتقسيم المعدّل (RSMA) في قنوات التّلاشي ذات الثّقْب المفتاحي. وتُعرف هذه البيئة ذات الرُّتبة المنخفضة بتأثيرها السلبي على أداء أنظمة الإدخال والإخراج الأحادي التّقليدية (SISO). وقد تمّ اشتقاق نتائج دقيقة وتقريبية لاحتمالية انقطاع الخدمة لتقنية (RSMA) بالنّسبة لقناة وصلة هابطة ثنائية المستخدمين، مع إلغاء التّداخُل المتتالي الكامل وغير الكامل. ويتمّ استخلاص الحلول المغلقة من خلال مقارنة حاصل ضرْب قناتَي تلاشٍ مستقلّتين تُحاكيان تأثير تقْب المفتاح. ولإثبات متانة تقنية الوصول المتعدّد بتقسيم المعدّل في مثل هذه البيئة، ندْرس أيضاً رتبة التّنوُّع ونكشف تأثير نقْص الرُّتبة النّاتج عن ثقْب المفتاح على موثوقية النّظام. وتُظهر نتائجنا أن التّقنية المذكورة تحتفظ بميزة أداء على الوصول المتعدّد غير المتعامد، خاصّةً مع عدم مثالية إلغاء التّداخل الثنائي أو انخفاض نسبة الإشارة إلى الضّجيج. وتؤكّد النّتائج العدّدية ومحاكاة مونت كارلو الصِّيغ النظرية، وتُبين أنّ تقنية الوصول المتعدّد بتقسيم المعدّل يمكنها مكافحة الآثار الضّارّة لقناة ثقْب المفتاح بشكلٍ أفضل من المخطّطات التّقليدية الأخرى. كذلك تؤكّد النّتائج امكانات تلك التّقنية لأنظمة الاتّصالات اللاسلكية المستقبلية الّتي تعمل في بيئات تلاشٍ شديد.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

ON THE RELIABILITY AND SPECTRAL EFFICIENCY OF MULTI-ANTENNA AF RELAY-AIDED NOMA NETWORKS

Hong-Nhu Nguyen[1], Mui Van Nguyen[2], Minh Xuan Pham[3]
and Sang-Quang Nguyen[3]

(Received: 27-Feb.-2026, Revised: 30-Apr.-2026, Accepted: 4-May-2026)

ABSTRACT

This paper investigates a downlink cooperative non-orthogonal multiple access (NOMA) system assisted by a multi-antenna amplify-and-forward (AF) relay. Unlike conventional single-antenna relay configurations, the considered framework jointly exploits relay diversity and direct transmission links between the base station (BS) and users. Under independent Nakagami-k fading channels, closed-form expressions for the outage probability (OP) and ergodic capacity (EC) of both users are derived for scenarios with and without direct BS-user links. The analytical formulation explicitly captures the effects of the number of relay antennas, fading severity and power allocation coefficients on system performance. For the ergodic capacity analysis, an exact integral representation combined with a Gaussian-Chebyshev quadrature approach is developed to efficiently evaluate the performance under the max-SINR selection criterion. The analytical results are verified through Monte-Carlo simulations and compared with orthogonal multiple access (OMA) benchmarks. Numerical results demonstrate that increasing the number of relay antennas significantly improves reliability due to enhanced spatial diversity. Moreover, the NOMA scheme achieves superior outage performance for the far user compared with OMA, while the ergodic capacity of the near user exhibits a pronounced gain in the moderate-to-high SNR regime. These findings confirm the effectiveness of multi-antenna cooperative relaying in improving both reliability and spectral efficiency.

KEYWORDS

Cooperative non-orthogonal multiple access (NOMA), Multi-antenna AF relaying, Spatial diversity, Nakagami-k fading, Outage probability, Ergodic capacity.

1. INTRODUCTION

Driven by the stringent requirements of next-generation wireless networks, non-orthogonal multiple access (NOMA) has emerged as a promising multiple access technique for fifth-generation (5G) and beyond systems [1]-[2]. In NOMA, multiple users are multiplexed in the power domain, where users experiencing poor channel conditions are allocated higher transmit power to guarantee fairness [3]. At the transmitter, superposition coding is employed to combine users' signals, while successive interference cancellation (SIC) is performed at the receivers to extract the desired information [4]. Owing to its flexible architecture, NOMA has also been extended to various emerging communication paradigms, including wireless-powered relaying and secure transmission frameworks [5].

To further enhance coverage and transmission reliability, cooperative NOMA (CNOMA) has been proposed, where relay nodes assist users with unfavorable channel conditions by forwarding the superimposed signals from the base station (BS) [6]-[7]. Relay-assisted transmission is particularly beneficial when the direct link between the BS and the far user suffers from severe path loss or shadowing effects. By introducing cooperative diversity, CNOMA systems can significantly improve reliability compared with non-cooperative NOMA architectures.

In recent years, the integration of NOMA with advanced enabling technologies has attracted increasing attention. For instance, the combination of NOMA and reconfigurable intelligent surfaces (RISs) has been investigated to improve physical layer security and reliability [8]. Active RIS-assisted dual-hop NOMA systems over Nakagami-κ fading channels were analyzed in [9], where closed-form outage and intercept probability expressions were derived to characterize both reliability and secrecy performance. Furthermore, RIS-assisted short-packet NOMA systems were studied in [10], highlighting

1. H.-N. Nguyen is with Faculty of Technology and Eng., Saigon Uni. (SGU), Ho Chi Minh City, Vietnam. Email: nhu.nh@sgu.edu.vn
2. M. V. Nguyen is with Gia Dinh Uni., Ho Chi Minh City, Vietnam. Email: muinv@giadinh.edu.vn
3. M. X. Pham and S.-Q. Nguyen (Corresponding Author) are with Posts and Telecommunications Institute of Technology (PTIT), Ho Chi Minh City, Vietnam. Emails: minhpx@ptit.edu.vn and sangnq@ptit.edu.vn

the impact of finite block-length transmission on secure communication. Beyond infrastructure-based scenarios, RIS-enabled device-to-device (D2D) NOMA frameworks with imperfect SIC were examined in [11], while partial NOMA-assisted backscatter communication systems were investigated in [12] to improve energy efficiency and spectrum utilization. These studies demonstrate the versatility of NOMA when combined with relay technologies, RIS, D2D communication and energy-efficient transmission mechanisms. More recently, several studies have further extended NOMA frameworks toward more practical and performance-oriented communication scenarios. For instance, short-packet NOMA systems have been investigated to support ultra-reliable low-latency communication (URLLC), where the interplay among latency, reliability and secrecy becomes critical under finite blocklength regimes [13]. In parallel, the integration of UAV-assisted communications with SWIPT has been explored to enhance system flexibility and energy efficiency, where joint optimization of power allocation, energy harvesting and UAV deployment plays a key role in improving overall network performance [14]. Moreover, active RIS-enhanced NOMA architectures have been proposed to dynamically reconfigure the wireless propagation environment, enabling both signal amplification and phase adaptation, thereby significantly improving outage performance, throughput and energy efficiency compared with conventional passive designs [15]. In addition, the incorporation of physical layer security and multi-antenna diversity techniques has been shown to effectively enhance robustness against fading and eavesdropping, providing improved secrecy performance in wireless communication systems [16].

These recent developments highlight a clear trend toward more realistic and performance-driven NOMA system designs. However, despite these advances, many existing works focus on integrating specific technologies rather than providing a unified analytical characterization that jointly captures multiple system components.

Despite these advancements, relay-assisted CNOMA systems remain a fundamental and practically relevant architecture due to their deployment simplicity and compatibility with existing networks. Early investigations primarily focused on single-antenna relay configurations [17]-[20], which are attractive for their low hardware complexity. However, single-antenna relays offer limited spatial diversity and interference-suppression capability, leading to performance degradation compared with multi-antenna counterparts [21]-[22]. By contrast, multi-antenna relays can exploit spatial diversity and array gain to significantly enhance transmission reliability and system robustness [23].

In addition to advanced technologies, such as RIS-assisted transmission and DF relaying, amplify-and forward (AF) relay architectures remain highly relevant due to their lower implementation complexity and analytical tractability. In particular, AF relaying avoids signal decoding at the relay, making it suitable for latency-sensitive and resource-constrained systems.

Therefore, the considered multi-antenna AF relay model serves as a canonical and analytically tractable framework for performance evaluation in cooperative NOMA systems. It enables a clear isolation of the gains introduced by cooperative relaying and spatial diversity, thereby providing a meaningful reference for comparison with more advanced, but structurally different, architectures.

Motivated by these benefits, multi-antenna relay-assisted CNOMA systems have been extensively studied under various configurations. A two-user CNOMA system employing a multi-antenna decode-and-forward (DF) relay was analyzed in [24], where outage probability and diversity order were derived in closed form. In [22], a multi-antenna two-way DF relay was considered to improve reliability. The impact of multi-antenna AF/DF relays on secrecy performance was examined in [25], revealing that increasing the number of antennas may introduce complex trade-offs between diversity and information leakage. Multi-antenna downlink NOMA systems were also investigated in [26], where power allocation and feedback overhead were jointly optimized.

Additionally, antenna-and-relay selection strategies were explored in [27]-[28] to reduce implementation complexity while maintaining diversity gains. More recently, full-duplex multi-antenna relaying has been incorporated into CNOMA systems to further enhance spectral efficiency [29]-[30].

Despite the extensive body of work on cooperative and RIS-assisted NOMA systems, existing studies have primarily focused on either decode-and-forward (DF) relaying strategies or RIS-enabled transmission frameworks. Recent works have also considered AF relay-assisted NOMA systems under different design objectives. For instance, the study in [31] investigates an AF relay-aided coordinated direct and relay transmission protocol to enhance energy efficiency and throughput in IoT networks

under imperfect SIC. Meanwhile, the work in [32] focuses on an AF MIMO two-way relay-assisted cognitive radio NOMA system with SWIPT, where secrecy performance is optimized *via* joint beamforming and power allocation.

In particular, multi-antenna DF relay-assisted NOMA works mainly emphasize outage performance, relay/antenna selection and diversity analysis under specific protocol settings [22], [24], while RIS-assisted NOMA approaches typically focus on beamforming design, achievable rate enhancement, SWIPT integration or secure transmission [8][9][10][11][12]. Compared with these studies, existing AF relay-based works are largely oriented toward protocol design or optimization objectives and comprehensive analytical characterizations remain limited.

However, to the best of our knowledge, the joint impact of multi-antenna amplify-and-forward (AF) relaying and direct transmission links on both reliability and spectral efficiency has not been fully characterized in a unified analytical framework. Most existing studies consider either relay-assisted transmission or direct links in isolation, which may not accurately reflect practical deployment scenarios.

To bridge this gap, this paper develops an analytically tractable model for a multi-antenna AF relay assisted NOMA system with MRC/MRT processing, where both direct BS-user links and relay-assisted links are jointly considered. Based on this unified framework, closed-form expressions for outage probability and tractable formulations for ergodic capacity are derived over Nakagami-κ fading channels, enabling a comprehensive characterization of both reliability and spectral efficiency.

The main contributions of this paper are summarized as follows:

- A cooperative NOMA system assisted by a multi-antenna AF relay is formulated, where both direct and relay-assisted transmission links are considered.
- Closed-form analytical expressions for the outage probability and ergodic capacity are derived over Nakagami-κ fading channels and verified *via* Monte-Carlo simulations.
- A comparative analysis is provided by considering OMA and NOMA transmission scenarios with and without direct links and different relay-antenna configurations. These baselines are selected to isolate the performance gains contributed by power-domain multiplexing, relay-assisted transmission and spatial diversity in a unified analytical framework.
- The impacts of transmit signal-to-noise ratio (SNR) and the number of relay antennas on system performance are thoroughly investigated, offering useful insights for practical system design.

Notation: Vectors are denoted by boldface letters, e.g., $\mathbf{x}$. The Frobenius norm is represented by $\|\cdot\|_F$ and $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and Hermitian transpose, respectively.

2. System Architecture and Signal Model

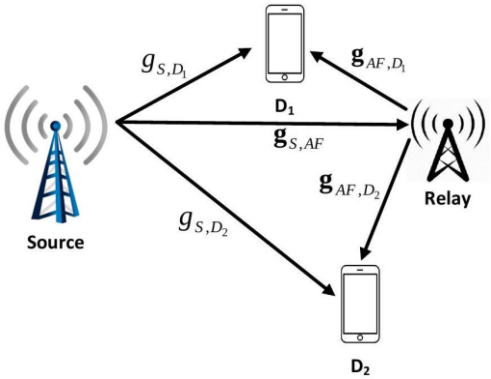


Figure 1. Illustration of the proposed cooperative NOMA architecture.

This section details the system architecture for a downlink cooperative non-orthogonal multiple access (NOMA) network. The layout comprises a single-antenna Source (S) communicating with a pair of single-antenna destinations, labeled as D_1 and D_2. This transmission is facilitated by a half-duplex amplify-and-forward (AF) relay node equipped with an array of K antennas, as depicted in Fig. 1. We assume that all communication links experience mutually independent Nakagami- κ fading. This channel model is adopted due to its flexibility in characterizing a wide range of fading environments,

including the standard Rayleigh fading case as a special scenario. Correspondingly, the channel power gains follow a Gamma distribution, uniquely defined by the severity parameter κ and the average channel gain λ. Additive white Gaussian noise (AWGN) with zero mean and variance σ^2 is assumed to impair all receiving nodes.

Let $\mathbf{g}_{S,AF} \in \mathbb{C}^{K\times 1}$ represent the channel vector spanning from the Source to the K-antenna relay, while $\mathbf{g}_{AF,D_i} \in \mathbb{C}^{K\times 1}$ denotes the link between the relay and destination $D_i (i \in \{1,2\})$. The direct line-of-sight (or non-line-of-sight) channel from the Source to D_i is defined by $g_{S,D_i} \in \mathbb{C}$. Given the Nakagami- κ environment, the respective channel power gains, defined as $|g_{S,D_i}|^2, \|\mathbf{g}_{S,AF}\|_F^2$ and $\|\mathbf{g}_{AF,D_i}\|_F^2$, adhere to Gamma distributions. These are parameterized by shape parameters $\kappa_{D_i}, \kappa_{AF}$ and κ_{AF,D_i}, alongside scale parameters (average powers) $\lambda_{D_i}, \lambda_{AF}$ and λ_{AF,D_i}. The noise entities at the relay and users are standard independent complex Gaussian variables, represented by $w_{AF}, w_{D_i} \sim \mathcal{CN}(0, \sigma^2)$.

During the primary transmission stage, the Source broadcasts a combined signal using superposition coding, encompassing two distinct messages s_1 and s_2 associated with the users. Assuming normalized symbol energies $\mathbb{E}[|s_1|^2] = \mathbb{E}[|s_2|^2] = 1$, the broadcast signal is formed as:

$$s_{tx} = \sum_{j=1}^{2} \sqrt{\alpha_j P_S} s_j, \tag{1}$$

where α_1 and α_2 act as the NOMA power allocation factors. To uphold the core NOMA methodology, these coefficients satisfy $\alpha_2 > \alpha_1$ and $\alpha_1 + \alpha_2 = 1$. Consequently, the signal observed at destination D_i is formulated as:

$$r_{D_i} = g_{S,D_i} s_{tx} + w_{D_i}, i \in \{1,2\}. \tag{2}$$

At the AF relay side, a maximum ratio combining (MRC) technique is leveraged. The corresponding receive weight vector is given by:

$$\mathbf{w}_{AF}^H = \frac{\mathbf{g}_{S,AF}^H}{\|\mathbf{g}_{S,AF}\|_F}$$

Therefore, the aggregated baseband signal arriving at the relay takes the form:

$$r_{AF} = \mathbf{w}_{AF}^H \mathbf{g}_{S,AF} s_{tx} + w_{AF}. \tag{3}$$

In the subsequent orthogonal phase, the relay scales and forwards its observed signal as $s_{AF} = \beta r_{AF}$. Here, the variable β denotes the relay amplification factor, strictly chosen to obey the relay's transmit power limit P_{AF}. This factor is computed as:

$$\beta = \sqrt{\frac{P_{AF}}{P_S \|\mathbf{g}_{S,AF}\|_F^2 + \sigma^2}}. \tag{4}$$

It is important to note that the adopted amplification factor corresponds to a variable-gain AF relaying scheme, where the relay gain dynamically adapts to the instantaneous channel state information (CSI) of the source-to-relay link. Compared with fixed-gain AF relaying, this approach effectively mitigates noise amplification and provides better performance in many practical scenarios. Therefore, the adopted model reflects a more realistic relaying operation while preserving the analytical tractability of the derived expressions.

For analytical convenience throughout this work, we presume uniform transmit power across the Source and the relay ($P_{AF} = P_S = P$).

At the relay, maximum ratio combining (MRC) is first applied to the received signal to maximize the received SNR across relay antennas. The combined signal is then amplified and forwarded using maximum ratio transmission (MRT), where the transmit beamforming vector is aligned with the corresponding relay-to-destination channel to enhance the end-to-end link quality.

It is important to note that the relay forwards the amplified version of the superimposed NOMA signal, i.e., the composite signal containing both s_1 and s_2, without performing signal separation or decoding. Therefore, the NOMA superposition structure is preserved across both the direct and relay-assisted links

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

and the separation of user signals is performed only at the receivers *via* successive interference cancellation (SIC).

For notational convenience, the MRT beamforming vector is expressed with respect to each destination link as $\mathbf{f}_i$, although the relay forwards the same composite signal without performing user-specific precoding.

Accordingly, no explicit multi-user precoding is performed at the relay and MRT is solely used to align the transmitted composite signal with the corresponding relay-to-destination channel, thereby enhancing the effective received SNR.

$$\mathbf{f}_i = \frac{\mathbf{g}_{AF,D_i}}{\left\|\mathbf{g}_{AF,D_i}\right\|_F}, i \in \{1,2\} \tag{5}$$

From a practical implementation perspective, the adopted MRC/MRT processing exploits all available relay antennas to enhance the array and diversity gains. Compared with simpler antenna selection or selection combining (SC) schemes, MRC/MRT requires the channel state information (CSI) of all relay antenna branches and the computation of receive/transmit weighting vectors, resulting in a higher processing burden as the number of relay antennas K increases. However, this complexity remains moderate, since only linear combining and beamforming operations are involved. In contrast, SC provides a lower-complexity alternative by selecting the strongest branch, at the expense of reduced spatial diversity gain. Therefore, MRC/MRT is adopted in this work to fully characterize the reliability and spectral-efficiency benefits of multi-antenna AF relaying.

Consequently, the forwarded signal captured by D_i from the relay node is given by:

$$r_{AF,D_i} = \mathbf{g}_{AF,D_i}^H \mathbf{f}_i s_{AF} + w_{AF,D_i}, i \in \{1,2\}. \tag{6}$$

To streamline the subsequent evaluations, we define the baseline transmit signal-to-noise ratio (SNR) as $\bar{\gamma} = \frac{P}{\sigma^2}$. Let us introduce the random variables $W_i = \left|g_{S,D_i}\right|^2, V = \left\|\mathbf{g}_{S,AF}\right\|_F^2$ and $Q_i = \left\|\mathbf{g}_{AF,D_i}\right\|_F^2$ to capture the channel power gains of the links $S \to D_i, S \to AF$ and $AF \to D_i$. Accordingly, the instantaneous SNRs over these respective paths are $\bar{\gamma}W_i, \bar{\gamma}V$ and $\bar{\gamma}Q_i$.

Due to the NOMA protocol, user D_2 considers the message s_1 as ambient noise. Thus, the instantaneous SINR at D_2 in the first phase is:

$$\Gamma_{D_2} = \frac{\alpha_2 \bar{\gamma} W_2}{\alpha_1 \bar{\gamma} W_2 + 1}. \tag{7}$$

Conversely, user D_1 first detects s_2 to perform successive interference cancellation (SIC). For analytical tractability, ideal SIC is assumed in this work, as commonly adopted in the literature to enable closed-form performance analysis. The SINR at D_1 for extracting s_2 is formulated as:

$$\Gamma_{D_1}^{s_2} = \frac{\alpha_2 \bar{\gamma} W_1}{\alpha_1 \bar{\gamma} W_1 + 1} \tag{8}$$

Accordingly, after successful interference cancellation, the remaining SNR for user D_1 to detect its intended message s_1 simplifies to:

$$\Gamma_{D_1} = \alpha_1 \bar{\gamma} W_1. \tag{9}$$

To account for practical imperfections, in realistic implementations, successive interference cancellation (SIC) may be imperfect due to channel-estimation errors, hardware limitations or residual interference. This effect can be modeled by introducing a residual interference factor $\rho \in [0,1]$, where $\rho = 0$ corresponds to ideal SIC and larger values of ρ indicate more severe residual interference.

In the presence of imperfect SIC, the effective SINR for the near user would be degraded due to the residual interference from the imperfect cancellation of s_2, leading to a higher outage probability and a reduction in ergodic capacity. Nevertheless, the overall performance trends with respect to key system parameters (e.g., the transmit SNR, the number of relay antennas K and the power allocation factors) remain consistent with those observed under the ideal SIC assumption.

Therefore, the ideal SIC assumption adopted in this work provides a useful analytical benchmark, while the above discussion offers practical insights into the impact of SIC imperfections without affecting the generality of the derived results. Regarding CSI, perfect channel knowledge is assumed in the analytical

derivations to maintain closed-form tractability. In practical systems, channel-estimation errors may lead to imperfect CSI and consequently perturb the effective channel gains used in the SINR expressions. Such uncertainty can be incorporated by modeling the estimated channel as the sum of the true channel and an estimation error term, e.g., $\hat{g} = g + e$, where e denotes the estimation error. This extension mainly modifies the effective channel statistics and SINR expressions, while the overall analytical framework remains applicable. A full closed-form treatment under imperfect CSI is left for future work due to the substantially increased analytical complexity. In particular, imperfect CSI typically reduces the effective channel gain and introduces additional uncertainty in the SINR expressions, leading to a performance degradation in both outage probability and ergodic capacity. Moving to the second phase, incorporating the amplification scalar and the MRT scheme, the end-to-end SINR at D_2 to retrieve s_2 *via* the relay path is calculated by:

$$\Gamma_{AF,D_2}^{s_2} = \frac{\alpha_2 \bar{\gamma}^2 V Q_2}{\alpha_1 \bar{\gamma}^2 V Q_2 + \bar{\gamma} V + \bar{\gamma} Q_2 + 1}. \tag{10}$$

Similarly, focusing on the $AF \to D_1$ cascade, the SINR at D_1 for the initial detection of s_2 and the subsequent SNR to decode its own data s_1, are represented, respectively, as:

$$\Gamma_{AF,D_1}^{s_2} = \frac{\alpha_2 \bar{\gamma}^2 V Q_1}{\alpha_1 \bar{\gamma}^2 V Q_1 + \bar{\gamma} V + \bar{\gamma} Q_1 + 1}, \tag{11}$$

$$\Gamma_{AF,D_1}^{s_1} = \frac{\alpha_1 \bar{\gamma}^2 V Q_1}{\bar{\gamma} V + \bar{\gamma} Q_1 + 1}. \tag{12}$$

Operating under a selection combining (SC) strategy, the ultimate effective SINRs available at destinations D_2 and D_1 are written as:

$$\Gamma_{D_2}^{eff} = \max\left(\Gamma_{D_2}, \Gamma_{AF,D_2}^{s_2}\right), \tag{13a}$$

$$\Gamma_{D_1}^{eff} = \max\left(\Gamma_{D_1}, \Gamma_{AF,D_1}^{s_1}\right). \tag{13b}$$

Although identical noise variance and representative large-scale fading parameters are adopted for clarity in the numerical evaluation, the proposed analytical framework remains sufficiently general to accommodate more realistic scenarios. In particular, heterogeneous settings with unequal path-loss and noise power across different links can be readily incorporated by appropriately adjusting the corresponding channel statistics. Therefore, such generalizations do not alter the fundamental structure of the derived expressions or the main performance insights.

3. Performance Metrics Analysis

3.1 Outage-probability Analysis

Given the nature of half-duplex relay networks, an entire communication cycle demands two orthogonal time slots. To guarantee specific quality-of-service (QoS) demands, the target SINR bounds for our nodes are mapped as:

$$\tau_{th,i} = 2^{2R_i} - 1, i \in \{1,2\},$$

where R_i indicates the pre-defined spectral efficiency threshold for D_i.

3.1.1 Outage Probability for Node D_2

Drawing upon standard literature [34]-[35], the cumulative distribution functions (CDFs) dictating the behavior of random variables W_i, V and Q_i are documented as:

$$F_{W_i}(x) = 1 - e^{-\eta_{D_i} x} \sum_{u=0}^{\kappa_{D_i}-1} \frac{\eta_{D_i}^u x^u}{u!}, i \in \{1,2\} \tag{14a}$$

$$F_V(x) = 1 - e^{-\eta_{AF} x} \sum_{u=0}^{\kappa_{AF}K-1} \frac{\eta_{AF}^u x^u}{u!} \tag{14b}$$

$$F_{Q_i}(x) = 1 - e^{-\eta_{AF,D_i} x} \sum_{u=0}^{\kappa_{AF,D_i}K-1} \frac{\eta_{AF,D_i}^u x^u}{u!}, i \in \{1,2\} \tag{14c}$$

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

The related probability density functions (PDFs) follow the shapes below:

$$f_{W_i}(x) = \frac{\eta_{D_i}^{\kappa_{D_i}}}{\Gamma(\kappa_{D_i})} x^{\kappa_{D_i}-1} e^{-\eta_{D_i} x}, i \in \{1,2\} \quad (15a)$$

$$f_V(x) = \frac{\eta_{AF}^{\kappa_{AF}K}}{\Gamma(\kappa_{AF}K)} x^{\kappa_{AF}K-1} e^{-\eta_{AF} x} \quad (15b)$$

$$f_{Q_i}(x) = \frac{\eta_{AF,D_i}^{\kappa_{AF,D_i}K}}{\Gamma(\kappa_{AF,D_i}K)} x^{\kappa_{AF,D_i}K-1} e^{-\eta_{AF,D_i} x}, i \in \{1,2\} \quad (15c)$$

Here, the constants are $\eta_{D_i} = \frac{\kappa_{D_i}}{\lambda_{D_i}}, \eta_{AF} = \frac{\kappa_{AF}}{\lambda_{AF}}$ and $\eta_{AF,D_i} = \frac{\kappa_{AF,D_i}}{\lambda_{AF,D_i}}$.

The outage likelihood for user D_2 evaluates the probability that its effective SINR falls short of the target τ_{th}:

$$O_2 = \Pr\left(\Gamma_{D_2}^{eff} < \tau_{th}\right) = \Pr\left(\max\left(\Gamma_{D_2}, \Gamma_{AF,D_2}^{S_2}\right) < \tau_{th}\right)$$
$$= \underbrace{\Pr\left(\Gamma_{D_2} < \tau_{th}\right)}_{\Psi_1} \underbrace{\Pr\left(\Gamma_{AF,D_2}^{S_2} < \tau_{th}\right)}_{\Psi_2} \quad (16)$$

The exact closed-form metric for D_2 's outage probability is derived as:

$$O_2 = \left(1 - e^{-\eta_{D_2}\vartheta} \sum_{u=0}^{\kappa_{D_2}-1} \frac{\eta_{D_2}^u \vartheta^u}{u!}\right) \times \left\{1 - 2 \sum_{v=0}^{\kappa_{AF}K-1} \sum_{j=0}^{v} \sum_{l=0}^{\kappa_{AF,D_2}K-1} \binom{v}{j} \binom{\kappa_{AF,D_2}K-1}{l} \right.$$
$$\times \frac{\eta_{AF}^v (\vartheta^2 + \vartheta\bar{\gamma}^{-1})^j \vartheta^{-(\kappa_{AF,D_2}K+v-j-l-1)} e^{-\vartheta(\eta_{AF,D_2}+\eta_{AF})}}{v!\,\Gamma(\kappa_{AF,D_2}K)}$$
$$\left. \times \eta_{AF,D_2}^{\kappa_{AF,D_2}K} \left(\frac{\eta_{AF}(\vartheta^2 + \vartheta\bar{\gamma}^{-1})}{\eta_{AF,D_2}}\right)^{\frac{l-j+1}{2}} \times K_{l-j+1}\left(2\sqrt{\eta_{AF}\eta_{AF,D_2}(\vartheta^2 + \vartheta\bar{\gamma}^{-1})}\right)\right\} \quad (17)$$

with the condition threshold defined as:

$$\vartheta = \frac{\tau_{th}}{\bar{\gamma}(\alpha_2 - \alpha_1\tau_{th})}$$

Proof 1. Refer to Appendix A.

3.1.2 Outage Probability for Node D_1

By a similar logic, D_1 's probability of communication failure is deduced as:

$$O_1 = \Pr\left(\Gamma_{D_1}^{eff} < \tau_{th}\right) = \Pr\left(\max\left(\Gamma_{D_1}, \Gamma_{AF,D_1}^{S_1}\right) < \tau_{th}\right) \quad (18)$$

Exploiting the statistical independence between the direct and relayed routes, we decompose the formula into:

$$O_1 = \underbrace{\Pr\left(\Gamma_{D_1} < \tau_{th}\right)}_{\Xi_1} \underbrace{\Pr\left(\Gamma_{AF,D_1}^{S_1} < \tau_{th}\right)}_{\Xi_2} \quad (19)$$

Computation of Ξ_1

Reflecting on $\Gamma_{D_1} = \alpha_1\bar{\gamma}W_1$, the failure state translates to:

$W_1 < \phi$, where we substitute $\phi = \frac{\tau_{th}}{\alpha_1\bar{\gamma}}$

Thus, the first term evaluates to:

$$\Xi_1 = F_{W_1}(\phi) = 1 - e^{-\eta_{D_1}\phi} \sum_{u=0}^{\kappa_{D_1}-1} \frac{\eta_{D_1}^u \phi^u}{u!} \quad (20)$$

Computation of Ξ_2

Referring back to (12), the boundary $\Gamma_{AF,D_1}^{s_1} < \tau_{th}$ algebraically shifts to:

$$VQ_1 < \phi(V + Q_1 + \bar{\gamma}^{-1})$$

A minor rearrangement isolates:

$$V < \frac{\phi(Q_1 + \bar{\gamma}^{-1})}{Q_1 - \phi}, \text{ under the premise } Q_1 > \phi.$$

As such, we formulate the probability integral:

$$\Xi_2 = 1 - \int_{\phi}^{\infty} f_{Q_1}(y) \left[1 - F_V\left(\frac{\phi(y + \bar{\gamma}^{-1})}{y - \phi}\right)\right] dy \tag{21}$$

Inserting the predefined mathematical distributions into (21) results in:

$$\Xi_2 = 1 - \sum_{p=0}^{\kappa_{AF}K-1} \frac{\eta_{AF,D_1}^{\kappa_{AF,D_1}K} \eta_{AF}^{p}}{p!\,\Gamma\left(\kappa_{AF,D_1}K\right)} \times \int_{\phi}^{\infty} y^{\kappa_{AF,D_1}K-1} e^{-\eta_{AF,D_1} y} e^{-\eta_{AF}\frac{\phi(y+\bar{\gamma}^{-1})}{y-\phi}} \left(\frac{\phi(y + \bar{\gamma}^{-1})}{y - \phi}\right)^{p} dy \tag{22}$$

Deploying variable substitution $z = y - \phi$ (thus $= z + \phi$), the equation morphs into:

$$\Xi_2 = 1 - \sum_{p=0}^{\kappa_{AF}K-1} \frac{\eta_{AF,D_1}^{\kappa_{AF,D_1}K} \eta_{AF}^{p} e^{-\phi(\eta_{AF,D_1}+\eta_{AF})}}{p!\,\Gamma\left(\kappa_{AF,D_1}K\right)} \times \int_{0}^{\infty} (z + \phi)^{\kappa_{AF,D_1}K-1} e^{-\eta_{AF,D_1} z} e^{-\frac{\eta_{AF}\phi(\phi+\bar{\gamma}^{-1})}{z}} \left(\phi + \frac{\phi^2 + \phi\bar{\gamma}^{-1}}{z}\right)^{p} dz \tag{23}$$

By utilizing the standard integral tables from ([33], Eq. (1.111), (3.471.9)), the integration can be entirely resolved. Skipping tedious algebra, the explicit closed-form result is:

$$\Xi_2 = 1 - 2 \sum_{p=0}^{\kappa_{AF}K-1} \sum_{w=0}^{p} \sum_{l=0}^{\kappa_{AF,D_1}K-1} \binom{p}{w} \binom{\kappa_{AF,D_1}K - 1}{l} \times \frac{\eta_{AF}^{p} \eta_{AF,D_1}^{\kappa_{AF,D_1}K} e^{-\phi(\eta_{AF,D_1}+\eta_{AF})}}{p!\,\Gamma\left(\kappa_{AF,D_1}K\right)} (\phi^2 + \phi\bar{\gamma}^{-1})^{w} \phi^{\kappa_{AF,D_1}K+p-w-l-1} \times \left(\frac{\eta_{AF}(\phi^2 + \phi\bar{\gamma}^{-1})}{\eta_{AF,D_1}}\right)^{\frac{l-w+1}{2}} K_{l-w+1}\left(2\sqrt{\eta_{AF}\eta_{AF,D_1}(\phi^2 + \phi\bar{\gamma}^{-1})}\right). \tag{24}$$

Ultimately, aggregating (20) and (24) into (19) yields the final outage equation for D_1.

3.1.3 High-SNR and Feasibility Insights

The derived outage expressions also provide useful insights into the high-SNR behavior and feasibility of the considered NOMA transmission. From the threshold variable $\vartheta = \frac{\tau_{\text{th}}}{\bar{\gamma}(\alpha_2 - \alpha_1\tau_{\text{th}})}$, it is clear that successful decoding of the far-user signal requires the feasibility condition $\alpha_2 > \alpha_1\tau_{th}$.

Equivalently, since $\alpha_1 + \alpha_2 = 1$, this condition can be written as $\alpha_2 > \tau_{\text{th}}/(1 + \tau_{\text{th}})$. If this condition is not satisfied, the required SINR threshold cannot be achieved even when the transmit SNR becomes large, which leads to an unavoidable outage floor for the far-user stream.

For the near user, the decoding of its own signal after successful SIC requires $\alpha_1 > 0$, while the prior decoding of the far-user message at D_1 follows the same feasibility condition $\alpha_2 > \alpha_1\tau_{th}$. Therefore, the power-allocation coefficients and target SINR threshold must be jointly selected to ensure meaningful outage performance for both users.

In the high-SNR regime, the outage probability generally follows the form $O_i \propto \bar{\gamma}^{-d_i}$, where d_i denotes the diversity order. Based on the small-argument behavior of the Gamma-distributed channel gains, the direct link contributes a diversity term governed by κ_{D_i}, whereas the relay-assisted path is mainly governed by the weaker hop between the $S \to AF$ and $AF \to D_i$ links, i.e., approximately $\min(\kappa_{AF}K, \kappa_{AF,D_i}K)$. Hence, under selection combining with both direct and relay-assisted links, the effective diversity order can be interpreted as:

$$d_i \approx \kappa_{D_i} + \min(\kappa_{AF}K, \kappa_{AF,D_i}K).$$

This observation explains why increasing the number of relay antennas K steepens the outage curves in the high-SNR region. Under the common fading-severity setting used in the simulations, this diversity gain increases approximately linearly with K, which is consistent with the trends observed in Figs. 2-3.

3.2 Ergodic-capacity Evaluation

We commence by formalizing the ergodic capacity function for the far node (D_2) :

$$C_{D_2} = \mathbb{E}\left\{\frac{1}{2}\log_2\left(1+\Gamma_{D_2}^{eff}\right)\right\} = \mathbb{E}\left\{\frac{1}{2}\log_2\left(1+\max\left(\Gamma_{D_2}, \Gamma_{AF,D_2}^{s_2}\right)\right)\right\}. \tag{25}$$

The analytical closed-form representation of D_2 's capacity is resolved as:

$$\begin{aligned} C_{D_2} = \frac{1}{2\ln 2}\Bigg\{ & \sum_{u=0}^{\kappa_{D_2}-1} \frac{\eta_{D_2}^u}{u!\,\bar{\gamma}^u}\left[\frac{1}{(\alpha_2+\alpha_1)^u} G_{1,2}^{2,1}\left(\frac{\eta_{D_2}}{\bar{\gamma}(\alpha_2+\alpha_1)}\middle| \begin{matrix} 1-u-1,- \\ 1-u-1,0 \end{matrix}\right)\right. \\ & \left. - \frac{1}{\alpha_1^u} G_{1,2}^{2,1}\left(\frac{\eta_{D_2}}{\bar{\gamma}\alpha_1^u}\middle| \begin{matrix} 1-u-1,- \\ 1-u-1,0 \end{matrix}\right)\right] + \frac{\pi^2}{2D} \sum_{v=0}^{\kappa_{AF}K-1} \sum_{j_1=0}^{v} \sum_{j_2=0}^{\kappa_{AF,D_2}K-1} \sum_{d=1}^{D} \binom{v}{j_1} \Bigg\} \\ & \times \binom{\kappa_{AF,D_2}K-1}{j_2} \frac{\eta_{AF,D_2}^{\kappa_{AF,D_2}} \eta_{AF}^v \sqrt{1-\Omega_d^2}}{v!\,\Gamma\left(\kappa_{AF,D_2}K\right)} \sec^2\left(\frac{\pi}{4}(\Omega_d+1)\right) \\ & \times \left(\frac{1}{\Upsilon(\Omega_d)+(\alpha_2+\alpha_1)^{-1}} - \frac{1}{\Upsilon(\Omega_d)+\alpha_1^{-1}}\right) \Theta(\Omega_d)^{j_1} \\ & \times \left(1 - e^{-\frac{\eta_{D_2}\Upsilon(\Omega_d)}{\bar{\gamma}}} \sum_{u=0}^{\kappa_{D_2}-1} \frac{\eta_{D_2}^u \Upsilon(\Omega_d)^u}{u!\,\bar{\gamma}^u}\right) \left(\frac{\Upsilon(\Omega_d)}{\bar{\gamma}}\right)^{\kappa_{AF,D_2}K+v-j_1-j_2-1} \\ & \times e^{-\frac{\Upsilon(\Omega_d)\left(\eta_{AF,D_2}+\eta_{AF}\right)}{\bar{\gamma}}} \left(\frac{\eta_{AF}\Theta(\Omega_d)}{\eta_{AF,D_2}}\right)^{\frac{j_2-j_1+1}{2}} K_{j_2-j_1+1}\left(2\sqrt{\eta_{AF}\eta_{AF,D_2}\Theta(\Omega_d)}\right)\Bigg\} \end{aligned} \tag{26}$$

where the roots for the Gaussian-Chebyshev approximation are:

$$\Omega_d = \cos\left(\frac{2d-1}{2D}\pi\right), d = 1, \ldots, D$$

and the auxiliary variables stand for:

$$\Upsilon(\Omega_d) = \tan\left(\frac{\pi(\Omega_d+1)}{4}\right), \Theta(\Omega_d) = \frac{\Upsilon(\Omega_d)}{\bar{\gamma}^2}(\Upsilon(\Omega_d)+1)$$

Proof 3. Refer to Appendix *B* for full derivations.

Progressing to the near node, the average capacity for D_1 is captured by:

$$\begin{aligned} C_{D_1} &= \mathbb{E}\left\{\frac{1}{2}\log_2\left(1+\Gamma_{D_1}^{eff}\right)\right\} = \mathbb{E}\{\frac{1}{2}\log_2(1+\underbrace{\max\left(\Gamma_{D_1}, \Gamma_{AF,D_1}^{s_1}\right)}_{U_{\text{var}}})\} \\ &= \frac{1}{2\ln 2}\int_0^{\infty} \frac{1}{1+y}\left[1-F_{U_{\text{var}}}(y)\right]dy \end{aligned} \tag{27}$$

The distribution function (CDF) of U_{var} is expanded mathematically as:

$$F_{U_{var}}(y) = \Pr\left(\max\left(\Gamma_{D_1}, \Gamma_{AF,D_1}^{S_1}\right) < y\right)$$
$$= 1 - e^{-\frac{\eta_{D_1}}{\alpha_1\bar{\gamma}}y} \sum_{t=0}^{\kappa_{D_1}-1} \frac{\eta_{D_1}^t y^t}{t!\,(\alpha_1\bar{\gamma})^t} \left[1 - 2 \sum_{p=0}^{\kappa_{AF}K-1} \sum_{w=0}^{p} \sum_{l=0}^{\kappa_{AF,D_1}K-1} \binom{p}{w} \right.$$
$$\times \binom{\kappa_{AF,D_1}K-1}{l} \frac{\eta_{AF,D_1}^{\kappa_{AF,D_1}K} \eta_{AF}^p e^{-\frac{y}{\alpha_1\bar{\gamma}}(\eta_{AF,D_1}+\eta_{AF})}}{p!\,\Gamma\left(\kappa_{AF,D_1K}\right)\left(\frac{y}{\alpha_1\bar{\gamma}} + \bar{\gamma}^{-1}\right)^w}$$
$$\times \left(1 - e^{-\frac{\eta_{D_1}}{\alpha_1\bar{\gamma}}y} \sum_{t=0}^{\kappa_{D_1}-1} \frac{\eta_{D_1}^t y^t}{t!\,(\alpha_1\bar{\gamma})^t}\right) \left(\frac{\eta_{AF}}{\eta_{AF,D_1}\alpha_1\bar{\gamma}}\left(\frac{y}{\alpha_1\bar{\gamma}} + \bar{\gamma}^{-1}\right)y\right)^{\frac{l-w+1}{2}}$$
$$\left. \times \left(\frac{y}{\alpha_1\bar{\gamma}}\right)^{\kappa_{AF,D_1}K+p-l-1} K_{l-w+1}\left(2\sqrt{\frac{\eta_{AF}\eta_{AF,D_1}}{\alpha_1\bar{\gamma}}\left(\frac{y}{\alpha_1\bar{\gamma}} + \bar{\gamma}^{-1}\right)y}\right)\right] \tag{28}$$

Applying (28) into (27), we partition C_{D_1} into two sub-integrals:

$$C_{D_1} = \frac{1}{2\ln 2}\left(I_{sub1} + I_{sub2}\right) \tag{29}$$

where the first piece acts as:

$$I_{sub1} = \sum_{t=0}^{\kappa_{D_1}-1} \frac{\eta_{D_1}^t}{t!\,(\alpha_1\bar{\gamma})^t} \int_0^\infty \frac{y^t e^{-\frac{\eta_{D_1}}{\alpha_1\bar{\gamma}}y}}{1+y}\,dy \tag{30}$$

Translating this through Meijer-G function properties yields:

$$I_{sub1} = \sum_{t=0}^{\kappa_{D_1}-1} \frac{\eta_{D_1}^t}{t!\,(\alpha_1\bar{\gamma})^t} G_{1,2}^{2,1}\left(\frac{\eta_{D_1}}{\alpha_1\bar{\gamma}}y \middle| \begin{matrix} 1-t-1, - \\ 1-t-1, 0 \end{matrix}\right) \tag{31}$$

Correspondingly, the second integral segment I_{sub2} emerges as:

$$I_{sub2} = 2 \sum_{p=0}^{\kappa_{AF}K-1} \sum_{w=0}^{p} \sum_{l=0}^{\kappa_{AF,D_1}K-1} \binom{p}{w} \binom{\kappa_{AF,D_1}K-1}{l} \frac{\eta_{AF,D_1}^{\kappa_{AF,D_1}K} \eta_{AF}^p}{p!\,\Gamma\left(\kappa_{AF,D_1}K\right)}$$
$$\times \int_0^\infty \frac{1}{1+y}\left(1 - e^{-\frac{\eta_{D_1}}{\alpha_1}y} \sum_{t=0}^{\kappa_{D_1}-1} \frac{\eta_{D_1}^t y^t}{t!\,(\alpha_1\bar{\gamma})^t}\right)\left(\frac{y}{\alpha_1\bar{\gamma}} + \bar{\gamma}^{-1}\right)^w$$
$$\times \left(\frac{y}{\alpha_1\bar{\gamma}}\right)^{\kappa_{AF,D_1}K+p-l-1} \left(\frac{\eta_{AF}}{\eta_{AF,D_1}}\left(\frac{y}{\alpha_1\bar{\gamma}} + \bar{\gamma}^{-1}\right)y\right)^{\frac{l-w+1}{2}}$$
$$\times e^{-\frac{y}{\alpha_1\bar{\gamma}}(\eta_{AF,D_1}+\eta_{AF})} K_{l-w+1}\left(2\sqrt{\eta_{AF}\eta_{AF,D_1}\left(\frac{y}{\alpha_1\bar{\gamma}} + \bar{\gamma}^{-1}\right)y}\right)dy \tag{32}$$

Adopting the Gaussian-Chebyshev node methodology to bypass intractability, I_{sub2} closely approximates to:

$$I_{sub2} \approx \frac{\pi^2}{2J} \sum_{p=0}^{\kappa_{AF}K-1} \sum_{w=0}^{p} \sum_{l=0}^{\kappa_{AF,D_1}K-1} \sum_{j=1}^{J} \binom{p}{w} \binom{\kappa_{AF,D_1}K-1}{l} \frac{\eta_{AF,D_1}^{\kappa_{AF,D_1}K} \eta_{AF}^p}{p!\,\Gamma\left(\kappa_{AF,D_1}K\right)}$$
$$\times \frac{\sqrt{1-\chi_j^2}}{1+\Upsilon(\chi_j)}\left(1 - e^{-\eta_{D_1}\Xi(\chi_j)} \sum_{t=0}^{\kappa_{D_1}-1} \frac{\eta_{D_1}^t \Xi(\chi_j)^t}{t!}\right)$$

$$\times \sec^2\left(\frac{\pi}{4}\left(\Upsilon(\chi_j)+1\right)\right)\left(\Xi(\chi_j)+\bar{\gamma}^{-1}\right)^{w} e^{-\Xi(\chi_j)\left(\eta_{AF,D_1}+\eta_{AF}\right)}$$

$$\times \Xi(\chi_j)^{\kappa_{AF,D_1}K+p-l-1}\left(\frac{\eta_{AF}\Pi(w)}{\eta_{AF,D_1}}\right)^{\frac{l-w+1}{2}} K_{l-w+1}\left(2\sqrt{\eta_{AF}\eta_{AF,D_1}\Pi(w)}\right) \quad (33)$$

in which $\chi_j = \cos\left(\frac{2j-1}{2J}\pi\right), \Upsilon(x) = \tan\left(\frac{\pi(x+1)}{4}\right), \Xi(x) = \frac{\Upsilon(x)}{\alpha_1\bar{\gamma}}$ and the term $\Pi(x) = \Xi(x)(\Xi(x) + \bar{\gamma}^{-1})$. Combining these solves C_{D_1} fully.

4. SIMULATION RESULTS

Within this section, we validate the derived analytical models through extensive numerical simulations. For simplicity across all evaluation scenarios, we assume a uniform fading severity environment by configuring the shape parameter as $\kappa = \kappa_{D_1} = \kappa_{D_2} = \kappa_{AF} = \kappa_{AF,D_1} = \kappa_{AF,D_2}$.

For clarity, this assumption is adopted to highlight the impact of key system parameters, such as the transmit SNR and the number of relay antennas. Nevertheless, in practical wireless environments, different communication links may experience heterogeneous fading conditions due to varying propagation characteristics (e.g., line-of-sight and non-line-of-sight components). It is important to emphasize that the proposed analytical framework does not rely on the assumption of identical fading parameters and the derived expressions remain valid when each link is characterized by a distinct Nakagami fading severity parameter κ. The use of a common κ value is therefore mainly for clarity and fair comparison purposes, while extending the analysis to heterogeneous fading scenarios is straightforward and left for future work.

The fundamental configuration metrics utilized for our network evaluation are summarized in Table 1. Furthermore, to guarantee a highly accurate evaluation of the integral approximations, the complexity terms for the Gauss-Chebyshev quadrature are fixed at $D = J = 100$.

Table 1. Summary of baseline simulation parameters.

Parameter Description	Notation	Assigned Value(s)
NOMA power allocation factors	$\{\alpha_1, \alpha_2\}$	{0.1, 0.9}
Nakagami fading-severity index	κ	2
Target SINR threshold	τ_{th}	2 (dB)
Number of antennas at the AF relay	K	{1,2,3}
Average channel-scale parameters	$\{\lambda_{D_1}, \lambda_{D_2}\}$	{1,1}
	λ_{AF}	1
	λ_{AF,D_1}	0.5
	λ_{AF,D_2}	0.9

We have verified that increasing the quadrature orders beyond these values (e.g., $J = 80,100,120$) results in negligible changes in the numerical results, thereby confirming the convergence behavior and accuracy of the adopted approximation. Specifically, the numerical computation of the ergodic capacity expressions requires $\mathcal{O}(D)$ and $\mathcal{O}(J)$ operations for the corresponding quadrature sums, while the remaining finite summations depend on the fading parameters and the number of relay antennas. Therefore, the proposed analytical evaluation remains computationally efficient and avoids the excessive runtime required by large-scale Monte-Carlo simulations. Specifically, the overall complexity scales linearly with the quadrature orders, i.e., $\mathcal{O}(D)$ or $\mathcal{O}(D + J)$ depending on the considered expressions.

Since the adopted quadrature orders are moderate (e.g., $J = 100$), the resulting computational burden remains low and tractable. Compared with conventional Monte-Carlo simulations, which typically require a large number of channel realizations to achieve statistical accuracy, the proposed analytical

approach significantly reduces computational cost while maintaining high accuracy. Therefore, the adopted approximation method provides an efficient and practical means for performance evaluation.

Fig. 2 depicts the system's outage probability as a function of the transmit SNR, $\bar{\gamma}$, across varying antenna-array sizes at the relay, K. Thanks to the core power-domain NOMA allocation rules, the far destination, D_2, inherently experiences superior outage reliability compared to the near destination, D_1. At lower $\bar{\gamma}$ regimes (between 0 and 10 dB), D_2 's outage rate stays practically at 10^0, reflecting a state dominated by noise and inter-user interference. However, as $\bar{\gamma}$ scales up, a sharp decline in outage probability is noticeable for both nodes, showcasing a much steeper decay at elevated SNR levels. Additionally, scaling the relay antenna count from $K = 1$ to $K = 3$ drastically mitigates outage events, verifying that augmented spatial diversity inherently fortifies the overall transmission robustness.

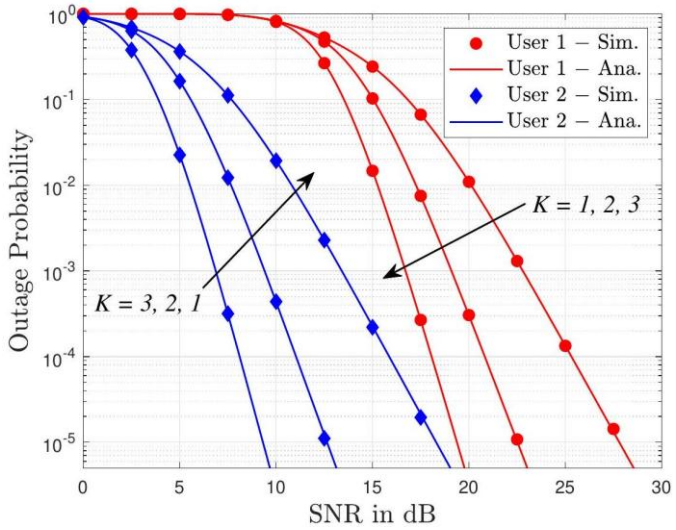


Figure 2. Outage probability *versus* transmitting SNR of two users.

Fig. 3 investigates the influence of the Nakagami fading severity index, κ, on the outage probability against $\bar{\gamma}$ when the relay is equipped with $K = 2$ antennas. Similar to previous observations, D_2 consistently preserves a lower outage profile than D_1 regardless of the transmission power. Naturally, higher $\bar{\gamma}$ translates to rapid drops in communication failures. Notably, boosting the shape parameter κ causes the performance curves to drop more precipitously, which implies a direct enhancement in the system's diversity order. Such trends perfectly align with the closed-form mathematics established earlier, confirming that higher κ values-indicative of less destructive fading conditions-yield substantially better reliability.

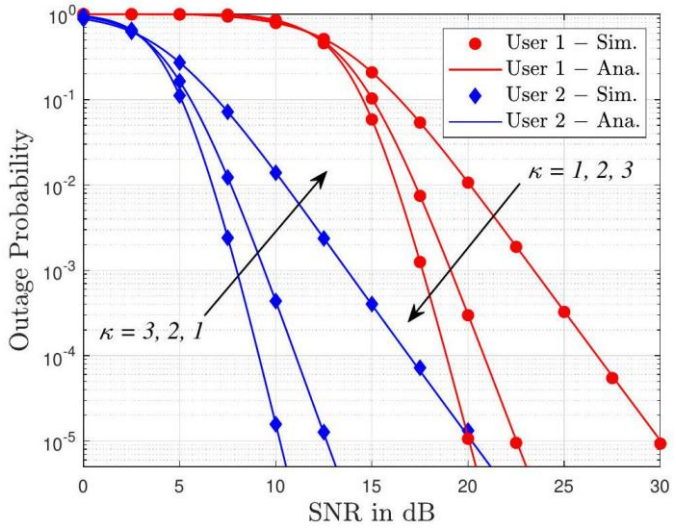


Figure 3. The outage probability *versus* SNR and different values of κ, with $K = 2$.

The impact of the power sharing factor, α_2, on the outage behavior is captured in Fig. 4, under a fixed $\bar{\gamma} = 15$ dB, $\kappa = 2$ and assorted K configurations. It is evident that dedicating a larger power fraction to

D_2 (increasing α_2) smoothly diminishes its probability of outage. Conversely, this exact increment starves D_1 of transmit power, consequently inflating its outage probability. This divergence perfectly encapsulates the fundamental power-domain compromises essential to NOMA frameworks. As expected, irrespective of the power split, expanding the relay's antenna array (K) continually upgrades both users' resilience by virtue of amplified spatial diversity gains.

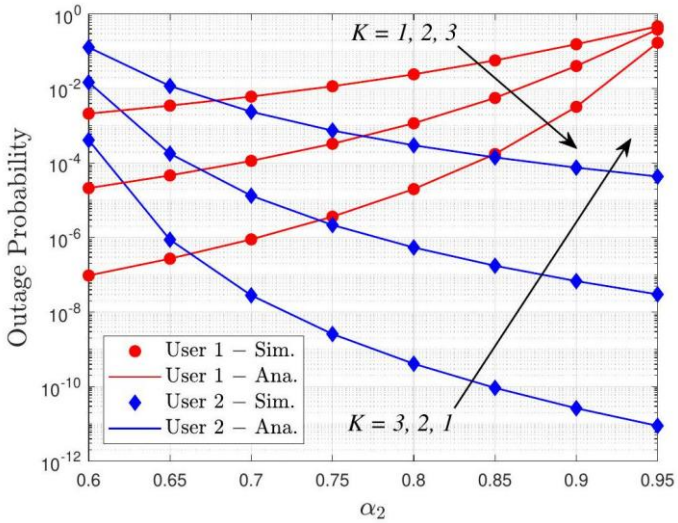


Figure 4. The outage probability *versus* α_2 with $\bar{\gamma} = 15(\text{ dB}), \kappa = 2$ and different values of K.

Fig. 5 contrasts the outage outcomes between network topologies that incorporate a direct $S \to D_i$ link *versus* those that strictly rely on the relay, evaluated at $\tau_{th} = 3$ dB and $K = 2$. Driven by the NOMA power-allocation hierarchy, D_2 constantly outshines D_1 in terms of reliability. Removing the direct transmission paths noticeably deteriorates the performance for both destinations, underscoring the critical advantage of leveraging the extra spatial diversity path. Moreover, transitioning the fading profile from $\kappa = 1$ to $\kappa = 2$ yields remarkable reliability upgrades in both topological setups, as milder fading intrinsically fortifies signal integrity.

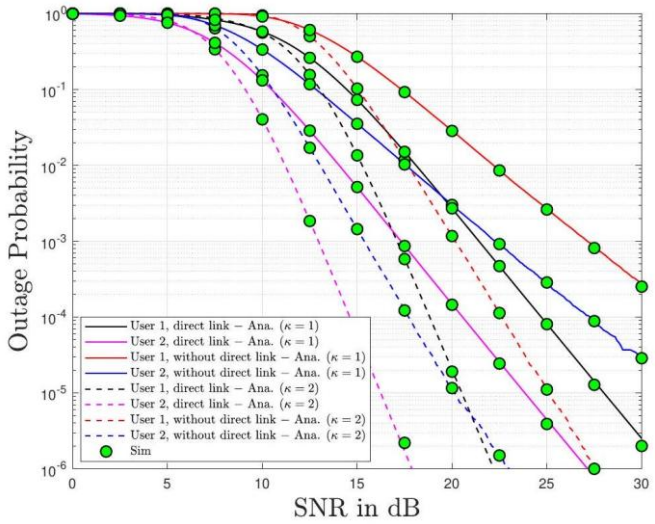


Figure 5. The outage probability with/without direct link *versus* $\tau_{\text{th}} = 3(\text{ dB}), \alpha_1 = 0.2$ and $K = 2$.

To provide meaningful and controlled baseline comparisons, the simulations evaluate the proposed NOMA scheme against OMA transmission, relay-assisted scenarios with and without direct links and different relay antenna configurations. These baselines are carefully selected to isolate the individual performance gains arising from power-domain multiplexing, cooperative relaying and spatial diversity. It is worth noting that the considered multi-antenna AF relaying framework serves as a fundamental benchmark for cooperative communication systems. While more advanced technologies, such as decode-and-forward relaying, reconfigurable intelligent surfaces and rate-splitting multiple access may

offer additional performance improvements, their integration typically requires substantially different system models and optimization strategies. Therefore, such comparisons are left for future work.

A direct reliability comparison between the proposed NOMA strategy and a conventional OMA baseline is portrayed in Fig. 6. The graphical data confirms that, for both $K = 1$ and $K = 2$ setups, the NOMA paradigm affords D_2 a noticeably lower outage probability than OMA. On the flip side, the near user (D_1) generally experiences better outage conditions under the OMA scheme than NOMA. Universally, elevating $\bar{\gamma}$ guarantees a swift plunge in the outage curves across all evaluated scenarios.

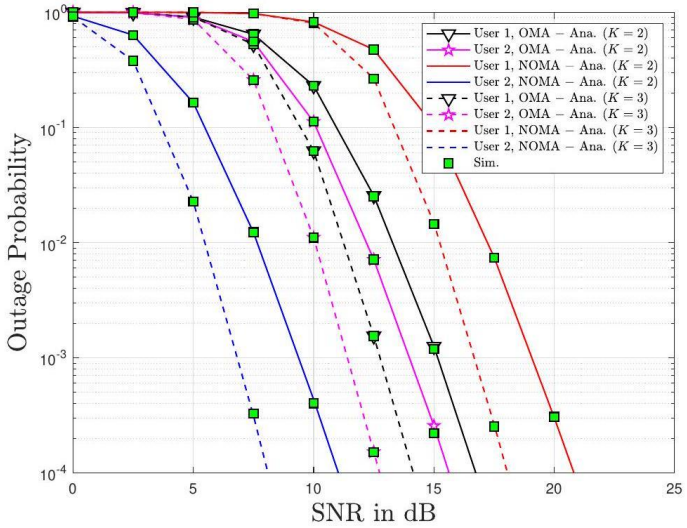


Figure 6. Comparison of outage probability between OMA and NOMA *versus* SNR, with $\kappa = 2$.

Fig. 7 plots the ergodic capacity trends against $\bar{\gamma}$ for varied relay antenna array dimensions ($K = 2,4,6$) while $\kappa = 2$. Analyzing the curves, D_2 's capacity climbs steadily within the low-to-medium $\bar{\gamma}$ bracket ($0 - 30$ dB), yet it begins to hit a ceiling at higher transmission powers. This saturation plateau occurs, because D_2 's throughput is heavily bottlenecked by inter-user interference and the fixed power allocation of NOMA. D_1, in stark contrast, enjoys an unabated capacity surge in the high- $\bar{\gamma}$ regime, thoroughly capitalizing on SIC mechanisms and amplified signal vigor. Equipping the relay with more antennas systematically elevates the ergodic capacity bounds for both nodes, thanks to superior spatial diversity. Nevertheless, for D_2, the added value of larger K arrays diminishes at high $\bar{\gamma}$, given that the environment shifts into an interference-limited state.

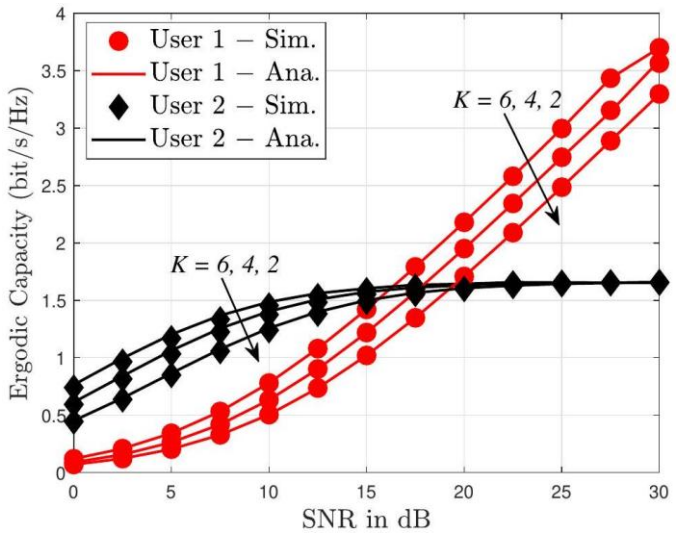


Figure 7. The ergodic capacity *versus* SNR and different values of K, with $\kappa = 2$.

Finally, Fig. 8 delineates the ergodic capacity as a function of the relay antenna count K when $\kappa = 2$. Throughout the evaluations, D_1 sustains a notably higher capacity output than D_2. Expanding K triggers substantial capacity boosts for D_1, a direct byproduct of the pronounced array and diversity gains

afforded by the multi-antenna relay. For D_2, however, the capacity only registers marginal improvements, effectively plateauing once K surpasses four antennas. This physical constraint aligns seamlessly with the derived analytical capacity equations, confirming that D_2 's upper bounds are dictated largely by NOMA power thresholds and persisting interference, thus capping the dividends paid by massive antenna expansions.

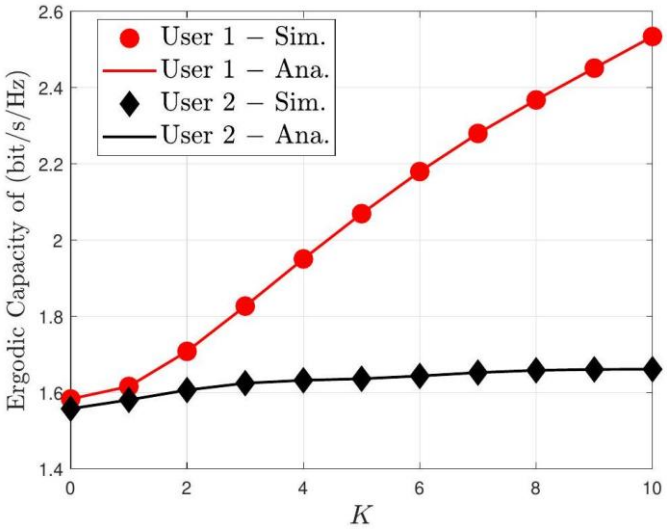


Figure 8. The ergodic capacity *versus* K, with $\kappa = 2$.

Although the analytical results are derived under the ideal SIC assumption, the impact of imperfect SIC can be interpreted as a degradation in the effective SINR due to residual interference, which results in a performance loss in terms of outage probability and ergodic capacity. However, the overall trends observed in the simulations remain unchanged.

5. CONCLUSIONS

This paper has analyzed a downlink cooperative NOMA system assisted by a multi-antenna amplify-and-forward relay over independent Nakagami- κ fading channels, where both direct and relay-assisted transmission links were considered. Closed-form expressions for the outage probability and ergodic capacity of both users were derived, providing an exact performance characterization.

The results show that employing multiple relay antennas significantly improves reliability by exploiting spatial diversity, leading to substantial outage reduction in the high-SNR region. The presence of direct BS-user links further enhances performance through selection combining. In addition, the proposed NOMA scheme outperforms OMA in terms of outage performance for the far user while achieving noticeable ergodic capacity gains at moderate-to-high SNR levels.

These findings confirm the effectiveness of multi-antenna AF relaying in enhancing reliability and spectral efficiency for cooperative NOMA systems. Future work may further consider more practical system impairments and asymmetries, including imperfect SIC, heterogeneous noise levels and link-specific path-loss conditions, to extend the proposed framework toward more realistic deployment scenarios. Future work may extend the current framework to incorporate more advanced communication paradigms, such as RIS-assisted transmission, DF relaying strategies and rate-splitting multiple access, to further enhance system performance under more complex network architectures.

APPENDICES

Appendix A - Proof of Proposition 1

Starting from the outage baseline, D_2 's probability of failure decomposes into:

$$O_2 = \underbrace{\Pr\left(\frac{\alpha_2 \bar{\gamma} W_2}{\alpha_1 \bar{\gamma} W_2 + 1} < \tau_{th}\right)}_{\Psi_1} \underbrace{\Pr\left(\Gamma_{AF,D_2}^{S_2} < \tau_{th}\right)}_{\Psi_2}. \quad (A.1)$$

Solving for Ψ_1

Analyzing the direct interference constraint:

$$\frac{\alpha_2 \bar{\gamma} W_2}{\alpha_1 \bar{\gamma} W_2 + 1} < \tau_{th}$$

we can isolate the channel gain requirement as:

$$W_2 < \vartheta,$$

assuming the boundary condition $\alpha_2 > \alpha_1\tau_{th}$ to prevent infinite thresholding, where:

$$\vartheta = \frac{\tau_{th}}{\bar{\gamma}(\alpha_2 - \alpha_1\tau_{th})}$$

This immediately confirms:

$$\Psi_1 = F_{W_2}(\vartheta) = 1 - e^{-\eta_{D_2}\vartheta}\sum_{u=0}^{\kappa_{D_2}-1}\frac{\eta_{D_2}^u\vartheta^u}{u!} \quad (A.2)$$

Solving for Ψ_2

Recalling the relayed SINR equation, the condition $\Gamma_{AF,D_2}^{S_2} < \tau_{\text{th}}$ is mathematically equivalent to bounding V:

$$VQ_2 < \vartheta(V + Q_2 + \bar{\gamma}^{-1}).$$

Restructuring provides:

$$V < \frac{\vartheta(Q_2 + \bar{\gamma}^{-1})}{Q_2 - \vartheta}, Q_2 > \vartheta.$$

Integrating over the PDF of Q_2 grants:

$$\Psi_2 = 1 - \int_{\vartheta}^{\infty} f_{Q_2}(y)\left[1 - F_V\left(\frac{\vartheta(y + \bar{\gamma}^{-1})}{y - \vartheta}\right)\right]dy \quad (A.3)$$

By expanding the probability definitions, Ψ_2 forms into:

$$\Psi_2 = 1 - \sum_{v=0}^{\kappa_{AF}K-1}\frac{\eta_{AF,D_2}^{\kappa_{AF,D_2}K}\eta_{AF}^v}{v!\,\Gamma(\kappa_{AF,D_2}K)}\int_{\vartheta}^{\infty} y^{\kappa_{AF,D_2}K-1}e^{-\eta_{AF,D_2}y}e^{-\eta_{AF}\frac{\vartheta(y+\bar{\gamma}^{-1})}{y-\vartheta}}\left(\frac{\vartheta(y + \bar{\gamma}^{-1})}{y - \vartheta}\right)^v dy \quad (A.4)$$

Applying a lateral shift $z = y - \vartheta$, the expression updates to:

$$\Psi_2 = 1 - \sum_{v=0}^{\kappa_{AF}K-1}\frac{\eta_{AF,D_2}^{\kappa_{AF,D_2}K}\eta_{AF}^v e^{-\vartheta\left(\eta_{AF,D_2}+\eta_{AF}\right)}}{v!\Gamma(\kappa_{AF,D_2}K)}\int_0^{\infty}(z+\vartheta)^{\kappa_{AF,D_2}K-1}e^{-\eta_{AF,D_2}z}e^{-\frac{\eta_{AF}\vartheta(\vartheta+\bar{\gamma}^{-1})}{z}}\left(\vartheta + \frac{\vartheta^2+\vartheta\bar{\gamma}^{-1}}{z}\right)^v dz.$$

Evaluating the integration parameters using Bessel transformation rules concludes the proof mapping identically to (17).

Appendix B - Proof of Proposition 2

Beginning with the integral definition of capacity for D_2 :

$$C_{D_2} = \mathbb{E}\left\{\frac{1}{2}\log_2\left(1 + \Gamma_{D_2}^{eff}\right)\right\} = \frac{1}{2\ln 2}\int_0^{\frac{\alpha_2}{\alpha_1}}\frac{1}{1+y}\left[1 - F_{X_{tmp}}\left(\frac{y}{\alpha_2 - \alpha_1 y}\right)\right]dy \quad (B.1)$$

Undergoing the variable conversion $z = \frac{y}{\alpha_2 - \alpha_1 y}$, (B.1) translates to:

$$C_{D_2} = \frac{1}{2\ln 2}\int_0^{\infty}\left(\frac{1}{z + (\alpha_2 + \alpha_1)^{-1}} - \frac{1}{z + \alpha_1^{-1}}\right)\left[1 - F_{X_{tmp}}(z)\right]dz \quad (B.2)$$

Linking back to the outage-probability derivations, $F_{X_{tmp}}(z)$ incorporates both direct and relayed elements, culminating in:

$$C_{D_2} = \frac{1}{2\ln 2}(\Delta_1 + \Delta_2) \quad (B.3)$$

Here, Δ_1 handles the un-relayed components:

$$\Delta_1 = \sum_{u=0}^{\kappa_{D_2}-1}\frac{\eta_{D_2}^u}{u!\,\bar{\gamma}^u}\int_0^{\infty}\left(\frac{1}{z + (\alpha_2 + \alpha_1)^{-1}} - \frac{1}{z + \alpha_1^{-1}}\right)z^u e^{-\frac{\eta_{D_2}}{\bar{\gamma}}z}dz \quad (B.4)$$

Transforming the exponential component using $e^{-bz} = G_{0,1}^{1,0}\left(bz\middle|_{0}^{-}\right)$, Δ_1 is computed strictly as:

$$\Delta_1 = \sum_{u=0}^{\kappa_{D_2}-1}\frac{\eta_{D_2}^u}{u!\,\bar{\gamma}^u}\left[\frac{1}{(\alpha_2 + \alpha_1)^u}G_{1,2}^{2,1}\left(\frac{\eta_{D_2}}{\bar{\gamma}(\alpha_2 + \alpha_1)}\middle|\begin{matrix}1-u-1,-\\1-u-1,0\end{matrix}\right) - \frac{1}{\alpha_1^u}G_{1,2}^{2,1}\left(\frac{\eta_{D_2}}{\bar{\gamma}\alpha_1^u}\middle|\begin{matrix}1-u-1,-\\1-u-1,0\end{matrix}\right)\right] \quad (B.5)$$

Meanwhile, Δ_2 dictates the relayed capacity segment, eventually resolved through numeric approximation (Gaussian-Chebyshev boundaries), completing the analysis.

REFERENCES

[1] Z. Ding et al., "A Survey on Non-orthogonal Multiple Access for 5G Networks: Research Challenges and Future Trends," IEEE Journal on Selected Areas in Communications, vol. 35, no. 10, pp. 2181-2195, 2017.

[2] L. Dai et al., "A Survey of Non-orthogonal Multiple Access for 5G," IEEE Communications Surveys & Tutorials, vol. 20, no. 3, pp. 2294-2323, 2018.

[3] Z. Wei et al., "A Survey of Downlink Non-orthogonal Multiple Access for 5G Wireless Communication Networks," arXiv preprint, arXiv: 1609.01856, 2016.

[4] S. R. Islam et al., "Power-domain Non-orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges," IEEE Comm. Surveys & Tutorials, vol. 19, no. 2, pp. 721-742, 2016.

[5] Q.-S. Nguyen, T. N. Nguyen and L.-T. Tu, "On the Security and Reliability Performance of SWIPT-enabled Full-duplex Relaying in the Non-orthogonal Multiple Access Networks," Journal of Information and Telecommunication, vol. 7, no. 4, pp. 462-476, 2023.

[6] M. Zeng et al., "Cooperative NOMA: State of the Art, Key Techniques and Open Challenges," IEEE Network, vol. 34, no. 5, pp. 205-211, 2020.

[7] Y. Yuan et al., "Joint Robust Beamforming and Power-splitting Ratio Design in SWIPT-based Cooperative NOMA Systems with CSI Uncertainty," IEEE Trans. on Vehicular Technology, vol. 68, no. 3, pp. 2386-2400, March 2019.

[8] A.-T. Le, T. D. Hieu, T. N. Nguyen, T.-L. Le, S. Q. Nguyen and M. Voznak, "Physical Layer Security Analysis for RIS-aided NOMA Systems with Non-colluding Eavesdroppers," Computer Communications, vol. 219, pp. 194-203, Apr. 2024.

[9] T. N. Nguyen, Q.-S. Nguyen, N. M. Quan, T. V. Chien, B. V. Minh and T.-L. Thuong, "Reliability and Security Analysis of Active RIS-assisted IoT NOMA Networks over Nakagami-*k* Fading Channels," IEEE Internet of Things Journal, Early Access, DOI:10.1109/JIOT.2025.3650486, 2026.

[10] N. Q. Sang et al., "Performance of RIS-secured Short-packet NOMA Systems with Discrete Phase-shifter to Protect Digital Content and Copyright against Untrusted User," IEEE Access, vol. 13, pp. 21580-21593, 2025.

[11] V.-D. Le et al., "Enabling D2D Transmission Mode of Reconfigurable Intelligent Surfaces Aided in Wireless NOMA System," Advances in Electrical and Electronic Eng., vol. 23, no. 1, 2025.

[12] T.-H. T. Pham et al., "Performance Analysis in D2D Partial NOMA-assisted Backscatter Communication," Advances in Electrical and Electronic Eng., vol. 23, no. 3, 2025.

[13] S.-Q. Nguyen et al., "Securing Short-packet Transmissions *via* Partial NOMA: Performance Analysis under Keyhole Fading," Vehicular Communications, vol. 58, p. 100999, 2026.

[14] A. Le-Thi et al., "Power Splitting-based SWIPT in UAV-aided NOMA Systems over Nakagami-*m* Fading: Performance Analysis and Optimization," Wireless Networks, vol. 32, pp. 315-330, 2026.

[15] M. Tran, M. Bui Vu and S. Nguyen, "On the Performance of Active RIS-enhanced NOMA Systems with Spectrum Sharing Mechanisms," Plos One, vol. 20, no. 11, p. e0336951, 2025.

[16] S.-Q. Nguyen et al., "Securing Wireless Communications with Energy Harvesting and Multi-antenna Diversity," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 11, no. 2, pp. 197-210, DOI: 10.5455/jjcit.71-1732244909, Jun. 2025.

[17] V. S. Nguyen, A. Le-Thi, V. D. Thuan, C.-B. Le, T. H. Nguyen and S.-Q. Nguyen, " Analysis of Ergodic Sum Rate in RSMA with Perfect and Imperfect SIC: A Multiple-antenna Selection Approach for Optimizing UAV Positioning," Physical Communication, vol. 72, p. 102741, 2025.

[18] N. H. Nhu et al., "Covert Communication Performance Evaluation in UAV-assisted Rate-splitting Multiple Access Systems," PLos One, vol. 20, no. 8, p. e0331013, 2025.

[19] Z. Yang, Z. Ding, Y. Wu and P. Fan, "Novel Relay Selection Strategies for Cooperative NOMA," IEEE Trans. on Vehicular Technology, vol. 66, no. 11, pp. 10114-10123, Nov. 2017.

[20] P. Xu, Z. Yang, Z. Ding and Z. Zhang, "Optimal Relay Selection Schemes for Cooperative NOMA," IEEE Trans. on Vehicular Technology, vol. 67, no. 8, pp. 7851-7855, Aug. 2018.

[21] M. Ashraf et al., "Energy Harvesting Non-orthogonal Multiple Access System with Multi-antenna Relay and Base Station," IEEE Access, vol. 5, pp. 17660-17670, 2017.

[22] L. Lv, Q. Ye, Z. Ding, Z. Li, N. Al-Dhahir and J. Chen, "On the Design of NOMA Assisted Multi-antenna Two-way Relay Systems," Proc. of the IEEE Int. Conf. on Communications (ICC), DOI: 10.1109/ICC40277.2020.9149110, Dublin, Ireland, 2020.

[23] X. Chen et al., "Exploiting Multiple-antenna Techniques for Non-orthogonal Multiple Access," IEEE Journal on Selected Areas in Communications, vol. 35, no. 10, pp. 2207-2220, Oct. 2017.

[24] Y. Cao et al., "Secrecy Analysis for Cooperative NOMA Networks with Multi-antenna Full-duplex Relay," IEEE Trans. on Communications, vol. 67, no. 8, pp. 5574-5587, Aug. 2019.

[25] N. Zaghdoud et al., "Secrecy Performance Analysis of Multi-antenna NOMA System with AF/DF Relaying under External and Internal Eavesdropping Scenarios," Proc. of the 2020 Int. Wireless Communications and Mobile Computing (IWCMC), pp. 1726-1732, Limassol, Cyprus, 2020.

[26] Z. Tang et al., "Reconsidering Design of Multi-antenna NOMA Systems with Limited Feedback," IEEE Trans. on Wireless Communications, vol. 19, no. 3, pp. 1519-1534, March. 2020.

[27] L. Lv, Q. Ye, Z. Ding, Z. Li, N. Al-Dhahir and J. Chen, "Multi-antenna Two-way Relay Based Cooperative NOMA," IEEE Trans. on Wireless Comm., vol. 19, no. 10, p. 64866503, 2020.

[28] T. A. Le and H. Y. Kong, "Energy Harvesting Relay-antenna Selection in Cooperative MIMO/NOMA Network over Rayleigh Fading," Wireless Net., vol. 26, no. 3, pp. 2075-2087, 2020.

[29] A. Hakimi, M. Mohammadi and Z. Mobini, "Outage Probability of Wireless-powered Multi-antenna Cooperative Spectrum Sharing Networks with Full-duplex and NOMA Transmissions," Proc. of the 2018 9th Int. Symposium on Telecommunications (IST), pp. 127-132, Tehran, Iran, 2018.

[30] Z. Mobini et al., "Full-duplex Multi-antenna Relay Assisted Cooperative Non-orthogonal Multiple Access," Proc. of the 2017 IEEE Global Communications Conference, pp. 1-7, Singapore, 2017.

[31] A. Jee and S. Prakriya, "Performance of Energy and Spectrally Efficient AF Relay-aided Incremental CDRT NOMA-based IoT Network with Imperfect SIC for Smart Cities," IEEE Internet of Things Journal, vol. 10, pp. 18766-18781, 2023.

[32] C. Hu, Q. Li, Q. Zhang and J. Qin, "Security Optimization for an AF MIMO Two-way Relay-assisted Cognitive Radio Non-orthogonal Multiple Access Networks with SWIPT," IEEE Trans. on Information Forensics And Security, vol. 17, pp. 1481-1496, 2022.

[33] I. S. Gradshteyn and I. M. Ryzhik, Tables of Integrals, Series and Products, 6th Edn., ISBN-10: 0-12-373637-4, New York: Academic Press, 2000.

[34] Q. Wang, J. Ge, Q. Li and Q. Bu, "Performance Analysis of NOMA for Multiple-antenna Relaying Networks with Energy Harvesting over Nakagami-*m* Fading Channels," Proc. of the 2017 IEEE/CIC Int. Conf. on Communications in China, pp. 1-5, Qingdao, China, Oct. 2017.

[35] C. C. Hung, C. T. Chiang, S. N. Lin and R. C. Wu, "Outage Capacity Analysis of TAS/MRC Systems over Arbitrary Nakagami-*m* Fading Channels," IEICE Trans. on Communications, vol. 93, no. 1, pp. 215-218, 2010.

ملخص البحث:

تبحث هذه الورقة في نظام الوصول المتعدّد غير المتعامِد التّعاوني (NOMA) للوصلة الهابطة، متعدّد الهوائيات، المدعوم بمرحل تضخيم وإعادة توجيه، على عكس تكوينات التّرحيل التّقليدية أحادية الهوائي.

يستغلّ الإطار المدروس تنوُّع التّرحيل وروابط الإرسال المباشر بين المحطّة الأرضية والمستخدمين. وفي ظلّ قنوات التلاشي المستقلّة، يتم اشتقاق صيغٍ مغلقة لاحتمالية انقطاع الخدمة والسّعة الإرجودية للمستخدمين في سيناريوهات وجود روابط مباشرة بين المحطّة الأساسية والمستخدمين وعدم وجودها.

ويوضّح النّموذج التّحليلي تأثير عدد هوائيات التّرحيل وشدّة التّلاشي ومعاملات تخصيص الطّاقة على أداء النّظام. ولتحليل السّعة الإرجودية، تمّ تطوير تمثيلٍ تكاملي دقيق مقترن بنهج تربيعي لتقييم الأداء بكفاءة. وقد تمّ التّحقُّق من النّتائج التّحليلية من خلال محاكاة مونت كارلو ومقارنتها بمعايير الوصول المتعدّد المتعامِد.

وأظهرت النّتائج العددية أنّ زيادة عدد هوائيات التّرحيل تحسّن الموثوقية بشكلٍ ملحوظ بفضل تعزيز التّنوُّع المكاني. علاوةً على ذلك، يحقّق نظام الوصول المتعدّد غير المتعامِد أداءً فائقاً في تقليل انقطاع الخدمة للمستخدِم البعيد مقارنةً بنظام الوصول المتعدّد المتعامِد، بينما تُظهر السّعة الإرجودية للمستخدِم القريب تحسُّناً ملحوظاً في نطاق نسبة الإشارة إلى الضّجيج المتوسّطة إلى العالية.

وتؤكّد هذه النّتائج فعالية تقنية التّرحيل التّعاوني متعدّد الهوائيات في تحسين كلٍّ من الموثوقية وكفاءة استخدام الطّيْف.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

ABPC-NET: A CAPSULE-GUIDED HYBRID FRAMEWORK FOR ROBUST ARABIC-TEXT CLASSIFICATION

Baqer M. Merzah[1] and Jafar Razmara[2]

(Received: 7-Mar.-2026, Revised: 18-Apr.-2026, 10-May-2026 and 16-May-2026, Accepted: 29-May-2026)

ABSTRACT

Arabic Text Classification (ATC) remains challenging due to the Arabic language's morphological richness and semantic complexity. This paper proposes ABPC-Net, a hybrid framework integrating a frozen Arabic Transformer encoder, a Bidirectional LSTM, parallel multi-scale CNN branches and a lightweight capsule-inspired vector projection head for hierarchical feature integration. Evaluated on the SANAD dataset and its subsets (AlArabiya, AlKhaleej and Akhbarona) over five independent runs, ABPC-Net achieves mean accuracies of 97.00±0.04%, 99.14±0.10%, 98.40±0.10% and 95.59±0.12%, respectively. Under identical experimental conditions, the proposed framework consistently outperforms re-implemented frozen and fully fine-tuned AraBERT and MARBERT baselines. Cross-dataset evaluation on BBC Arabic and CNN Arabic further provides evidence of intra-domain transferability and rapid few-shot adaptability across Arabic news sources. The reported results are scoped to Modern Standard Arabic news classification.

KEYWORDS

Arabic text classification, Deep learning, Transformer models, Capsule networks, Natural-language processing (NLP).

1. INTRODUCTION

Arabic Text Classification (ATC) is a fundamental problem in the field of Natural Language Processing (NLP), as it allows a wide range of applications, from information retrieval [1,2] to sentiment analysis, fake news detection [3], among others [4]. Though deep learning (DL) has achieved significant results in this field, ATC faces a unique set of challenges due to the complex morphology of the Arabic language, as well as the scarcity of large datasets compared to the English language [5]. This has led to a significant amount of scholarly work in the quest to create a robust model that accommodates the complexities of the Arabic language.

Moreover, DL models are challenging to leverage in the Arabic language, because the natural characteristics of the language differentiate it from Indo-European languages. Arabic has a rich root and pattern morphology, where a single trilateral root such as ك- ت- ب (k-t-b) can generate many possible semantic derivatives, such as 'katib' (كاتب - writer), 'maktaba' (مكتبة - library) and 'maktub' (مكتوب - written) [6]. Given this morphological mix, both the dimensionality and sparsity of the feature space are relatively high. The language heavily uses agglutination, where prepositions, conjunctions and pronouns are merged into the word stem. A notable example provided is the token (فسيكفيكهم- fasayakfeekahum), which is broken down into syntactic units: 'fa' (then) + 'sa' (will) + 'yakfi' (suffice) + 'ka' (you) + 'hum' (them). This hybridized morphology is not compatible with normal tokenizers [7, 8]. Additionally, the widespread omission of diacritics in Modern Standard Arabic causes orthographic confusion (homographs). For instance, the unvoweled word ذهب (dhb) can mean 'gold' or 'went' depending on the context [9]. The complexities of these features make shallow models inadequate, as they require architectures that can simultaneously consider local morphological features, sequence context and hierarchical semantic structures.

Previous studies on ATC have utilized common ML approaches and simple DL frameworks, such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks [10]. Elnagar et al. [5] made significant contributions to the field of ATC. These contributions comprise developing the SANAD dataset of Arabic news articles to establish a comprehensive benchmark. They

1. B. M. Merzah is with Department of Computer Science, Faculty of Education, Uni. of Kufa, Iraq. Email: baqirm.merzah@uokufa.edu.iq
2. J. Razmara is with Department of Computer Science, Faculty of Mathematics, Statistics and Computer Science, University of Tabriz, Tabriz, Iran. Email: razmara@tabrizu.ac.ir

additionally proposed multiple DL architectures, among which Attention-GRU was proposed for state-of-the-art (SOTA) performance. Modern investigations build upon these baseline methods. They propose hybrid architectures; for example, Inception-CNN in combination with LSTM [4]. These models provide additional evidence about the performance of hybrid neural architectures on various textual attributes.

Existing SOTA methods heavily rely on large pre-trained Transformer models (e.g., AraBERT), which are able to learn to capture deep contextual relationships between words. Moreover, despite these advancements, an important research gap is still to be filled. Still, the prevailing model consists of fitting a basic classification head (e.g., a single dense layer) to the output of the Transformer. Although this step is computationally efficient, it may be a bottleneck, since it cannot fully leverage the rich, high-dimensional representations from the Transformer and it cannot fully characterize the complex hierarchical relationships between the extracted features. The novelty of ABPC-Net lies in addressing this gap through three deliberate and complementary design decisions. First, a BiLSTM layer re-encodes transformer sequence outputs to capture long-range sequential dependencies with directional awareness suited to Arabic morphological structure. Second, a parallel multi-scale CNN module with kernel sizes of 2, 3 and 4 performs explicit n-gram feature extraction at multiple granularities, a design specifically motivated by Arabic agglutination. Third, a Capsule-inspired Vector Projection Head replaces the conventional scalar softmax classifier with a vector-based encoding mechanism that preserves multi-dimensional feature relationships, enabling richer integration of multi-scale representations. Unlike recent transformer-based models, such as Tasneef [11], CLGNet [12] and ABTM [13], which either attach simple dense layers or augment transformers with frequency-based features, ABPC-Net explicitly encodes hierarchical spatial relationships among contextualized features. Though a Transformer may decide what words are relevant in context, this feature-based approach may still not explicitly model how the contextualized features group together to form higher-level abstract concepts, something important for nuanced classification. To cope with this limitation, we explore using more sophisticated downstream architectures able to better interpret the rich representations generated by Transformers.

This work presents a structured downstream architectural design for transformer-based ATC and provides systematic empirical evidence that such a design can extract substantially more value from a frozen Arabic transformer encoder than full fine-tuning with a shallow classification head. The contribution is therefore primarily architectural and empirical in nature, supported by controlled experimental analysis. The main contributions of this paper are:

- Structured Downstream Architecture: We propose ABPC-Net, a structured downstream pipeline combining BiLSTM sequential re-encoding, parallel multi-scale CNN branches (kernel sizes 2, 3 and 4) and a lightweight capsule-inspired vector projection head atop a frozen Arabic transformer encoder. Their systematic integration and controlled evaluation provide new empirical insights for Arabic text classification.
- Mechanistic Analysis of Capsule-inspired Projection: We provide an ablation-driven analysis showing that the capsule-inspired projection behaves as a relational feature fusion mechanism the effectiveness of which depends on input structural richness. In particular, it degrades performance without BiLSTM, but consistently improves performance when preceded by sequential encoding, offering practical design insights.
- Comprehensive Empirical Validation: Experimental evaluation of SANAD is conducted extensively, along with source-specific performance analysis (Akhbarona, AlArabiya, AlKhaleej).
- Cross-dataset Generalization: We evaluate ABPC-Net on BBC Arabic and CNN Arabic through zero-shot transfer and few-shot domain adaptation protocols, characterizing the model's intradomain transferability across Arabic news sources.

Our results demonstrate that our model performs well in multiple news contexts and that we include a qualitative error analysis to show how the model behaves across different news domains. Our results show that the proposed hybrid architectures provide strong empirical performance, illustrated by the Capsule Networks' performance of the model, which means that it performs significantly better than the baseline. This work contributes a structured hybrid framework for Arabic news classification that demonstrates strong empirical performance on the SANAD benchmark and offers a methodological foundation for exploring multi-component architectures in Arabic NLP tasks. The remainder of this

paper is organized as follows. Related work is described in Section 2. The ABPC-Net architecture, along with the experimental settings and hyper-parameters, is described in Section 3. The results and discussion, including the ablation analysis, error analysis and limitations, are presented in Section 4. Finally, Section 5 concludes the paper and outlines future directions.

2. RELATED WORK

Arabic-text classification (ATC) has attracted tremendous emphasis in recent years and various tools were explored, such as classical machine learning (ML), especially classical approaches and advanced DL architectures. In this section, we present a review of how these techniques evolved. Among advanced ATC approaches, standard classical ML methods were widely applied, often used together with several feature-engineering techniques, such as TF-IDF and Bag-of-Words [14], including shallow-learning approaches for tagging Arabic news articles [15]. The field was transformed by DL. In the earliest DL works, it generally limited the process to processing base architectures (CNNs) to extract local features and recurrent neural networks (RNNs) (Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs)) for capturing sequential dependencies in text [16]. Alsaleh and Larabi-Marie-Sainte presented a hybrid model [17], on a CNN and Genetic Algorithm (GA) of ATC, representing the trend. However, their GA-CNN model, based on GloVe as the word embedding tool obtained relatively high accuracy on Moroccan and Saudi Newspaper Article Datasets, though it stated that the training time was prolonged.

In recognition of the fact that, while CNNs and RNNs fail to adequately capture the complexities involved in Arabic-text processing, subsequent research has focused on the development of more advanced hybrid architectures. This research direction, of which the present work is a part, focuses on the synergistic integration of the capabilities offered by different architectures. For example, in research that focused on the exploration of supervised-learning strategies, Ameur et al. [18] integrated the capabilities offered by different word embeddings, including static, dynamic and fine-tuned word embeddings, into RNNs and CNNs. The results obtained by the model, especially the integration of CNNs with Bidirectional Gated Recurrent Units (BiGRUs), indicated a significant level of effectiveness, as the F-score increased by as much as 98.61%. In similar research, Jamaleddyn et al. [12] proposed a novel multi-channel deep-learning model known as CLGNet, which integrates the capabilities offered by CNN, long short-term memory networks and Gated Recurrent Units. The results obtained by the model, following the extensive pre-processing and SMOTE-based balancing of the CNN, BBC and OSAC datasets, indicated a significant level of performance, as the model performed better than the capabilities offered by different deep-learning architectures.

The large-scale publicly available dataset SANAD was introduced by Elnagar et al. [5], which was an important milestone in the field. Their pioneering work has given an invaluable resource to the task and set a very high benchmark for many types of DL models, such as various hybrid CNN-RNN architectures and attention-based paradigms. They found that models incorporating attention, such as Attention-GRU, can achieve SOTA performance, thus raising the benchmark of the task. In recent times, Alnagi et al. [4] proposed a hybrid model by incorporating the Inception-CNN and LSTM layers. This further emphasizes the adoption of various neural architectures for the processing of textual features on multiple scales. Accuracies of 92% and 96% were reported for the SANAD and AlKhaleej datasets, respectively, by utilizing complex variants of the CNN algorithm.

Jalil and Aliwy [19] advanced a novel hybrid CNN-BiLSTM architecture to facilitate taskful workloads, like topic classification, sentiment analysis, emotion recognition and sarcasm detection. Combining convolutional layers for local feature extraction with LSTM units that bidirectionally describe and capture the contextual dependencies, the model performs well on embedding spatial and sequential representations. In general, their experimental results showed good performance on topic classification (97.58%) and sarcasm handling (97%), sentiment analysis (86%) and emotion recognition (81.6%). This observation demonstrates the ability of hybrid DL architectures to model much greater language complexity and variability within the context of Arabic social-media content that faces multiple challenges, from the limited length of text to informal-language use to implicit and contextual semantic cues in the field. Novel hybrid-based paradigms have further developed the state-of-the-art by exploiting rich context embeddings to their limit.

Hossain et al. [13] proposed a hybrid model known as Attention-based Transformer Model (ABTM),

consisting mainly of deep contextual data with traditional statistical features (e.g., TF-IDF and Bag-of-Words). The present study has delivered significant improvements in performance, realizing full and best-in-class performance of 97.69% accuracy on their Arabic news dataset. This is a clear indication of the growing realization of the effectiveness of combining context features with other architectural components. Along with these architectural advancements, other research has also focused on the importance of feature representation and hyper-parameter optimization. To cite an example, B. Al-onazi et al. [20] presented a hybrid model referred to as the CRNN model, which combines CNN and RNN architectures. However, the novelty in their model is the application of the Crow Search Algorithm for hyper-parameter optimization of their model. Even though their model has clearly demonstrated the importance of hyper-parameter optimization in improving model performance, it still relies on traditional features, such as TF-IDF, which do not fully exploit the power of deep context features in natural languages.

Recent SOTA mostly converges with large pre-trained Transformer models, such as AraBERT, GigaBERT and MARBERT, to model the subtle semantic structures. Yet, as noted above, a well-known limitation is that many of the applications embed a simple, shallow classifier head on the Transformer that does not take advantage of its rich, multi-layered representations. And recent works, such as the work by Hossain et al. [13], have attempted to augment Transformers with classical statistical features (TF-IDF, BoW) and these approaches continue to employ frequency-based representations that lack structural depth. In contrast, ABPC-Net introduces a structured downstream pipeline, BiLSTM for sequential re-encoding, parallel CNNs for multi-granularity local-feature extraction and a capsule-inspired vector projection for hierarchical feature integration, that qualitatively differs from the shallow-classification strategies employed by Tasneef [11], CLGNet [12] and ABTM [13]. Unlike these approaches, ABPC-Net explicitly encodes spatial and hierarchical relationships among contextualized features, preserving part-whole relationships characteristic of Arabic morphology for more robust and granular classification. Although capsule-based models have been explored in text classification, their behavior within hybrid transformer-based pipelines for Arabic-text classification has received limited systematic analysis. The present work contributes to filling this gap by providing a controlled ablation-driven analysis of how capsule-inspired projections interact with sequential and convolutional components in the Arabic-text classification setting.

3. MATERIALS AND METHODS

In this section, the proposed ABPC-Net model for the classification of Arabic news articles will be delineated.

The methodology is based on the development of a hybrid deep-learning model, which combines the power of the pre-trained transformer model, recurrent neural network and parallel capsule network in a complementary fashion. There are four main steps in the proposed methodology: (1) Data Source, which describes the SANAD dataset; (2) Data Pre-processing, which describes the pre-processing steps for the data; (3) Model Architecture, which describes the ABPC-Net model; and (4) Training and Experimental Setup, which describes the settings used for the model's training process. A block diagram for the ABPC-Net model is presented in Figure 1.

3.1 Dataset

The SANAD dataset [5] is a large-scale Arabic news corpus used for single-label text classification. Even though the raw dataset consists of around 200,000 articles, their authors constructed a refined and balanced sub-set to minimize class imbalance and remove noisy or super brief texts. In this study, the pre-processed version was used based on their established benchmark. Consequently, the dataset used for our experiment consists of 110,900 high-quality articles distributed across the Al-Arabiya, AlKhaleej and Akhbarona portals: SANAD covers seven topical categories of Culture, Finance, Medicine, Politics, Religion, Sports and Technology. The dataset itself is balanced at the category level within each source-specific sub-set of these pieces, providing equal opportunities to evaluate classification models. In addition, SANAD is further divided over training and test partitions that are used in our demonstrations. The scale, linguistic consistency (MSA) and categorical diversity of the dataset, is significant enough to serve as a benchmark for DL-based ATC systems. The category-wise distribution of SANAD for its three source-dependent classes is shown in Table 1.

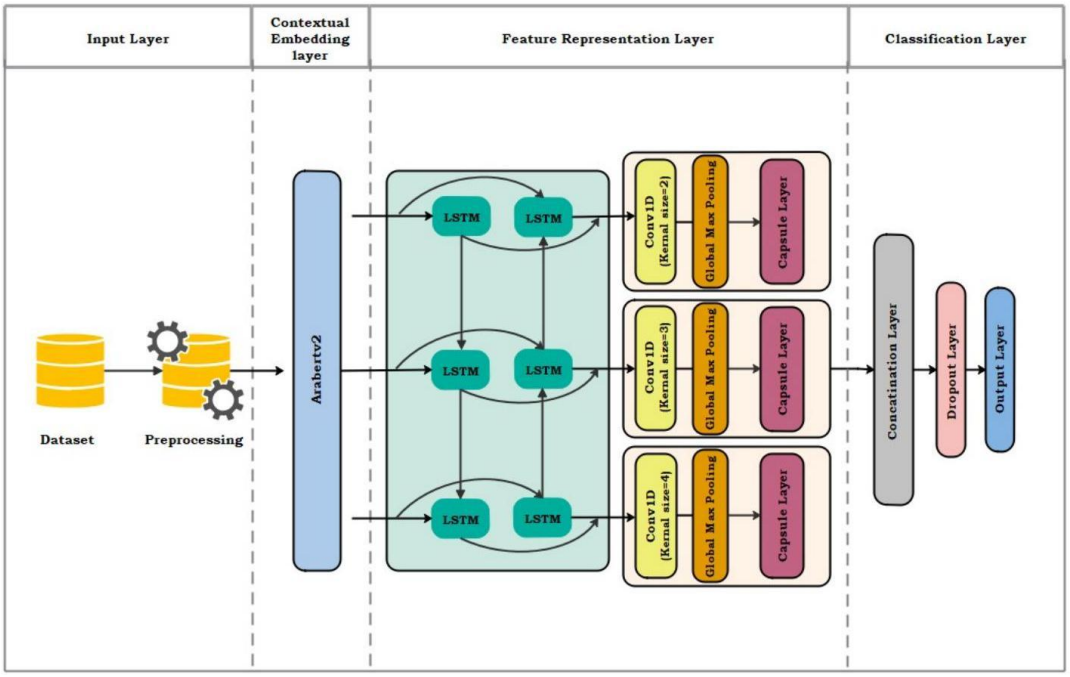


Figure 1. The ABPC-Net architecture, illustrating the sequential flow from frozen AraBERT embeddings through BiLSTM re-encoding, parallel CNN-capsule feature extraction and final classification *via* margin loss.

In particular, Al-Arabiya has five categories; Al-Khaleej and Akhbarona have all seven categories. We used the standard train-test split from the SANAD dataset, but 10% of the training dataset was allocated randomly for validation purposes. This internal validation split was used just to check how the model was performing during training and for triggering the Early Stopping mechanism. Thus, during all stages of training and hyper-parameter tuning, the official test set was solely unseen, so that at the end of testing the final evaluation was entirely unbiased.

Table 1. A balanced sub-set of SANAD articles and category counts per dataset.

Source	Categories	Training	Testing	Total	Per Category
alarabiya.net	5	16,650	1,850	18,500	3,700
alkhaleej.ae	7	40,950	4,550	45,500	6,500
akhbarona.com	7	42,210	4,690	46,900	6,700

3.2 Data Pre-processing

The text data contains noise, such as punctuation, numbers, non-Arabic scripts and diacritics, known as Tashkeel, which can negatively impact the performance of the model. Before feeding the text data into the model, a thorough and carefully tailored pre-processing plan for Arabic language is used. The major steps, which are explained in detail, are as follows:

- Normalization: Due to the fact that digital texts feature orthographic inconsistencies, a very indepth normalization is carried out. It is a crucial step, especially when the model is going to depend mainly on the semantics rather than the spelling. Among other things, this means removing all Tashkeel and Tatweel (character elongation) and unifying different forms of the Hamza (إ،أ،آ) to one form (ا). Moreover, 'Ta marbuta' (ة) is converted into 'Ha' (ه) and 'Alef maksura' (ى) is converted into 'Ya' (ي).
- Noise Removal: Repetitive and non-informative pieces are taken out of the text. This means to remove all punctuation marks, numerical digits and any Latin characters.
- Tokenization: The text after it is cleaned is broken down into single words using the Farasa

package, which is adjusted and optimized for Arabic text.

- Stop-word Removal: Concentrating on the functionality only, typical Arabic stop words, such as (e.g., من, في, على), which hardly provide any semantic value for classification, are removed using a pre-defined list from NLTK's Arabic corpus.

Such a strict and detailed pre-processing is our assurance that the data fed into our model network is cleansed, standardized and oriented towards semantically meaningful content.

3.3 Hybrid Model Architecture

Our proposed model is an end-to-end DL model designed to recognize intricate linguistic patterns. It combines the contextual capabilities of the transformer model, the sequential capabilities of the recurrent neural network and the hierarchical feature detection of the capsule network. Figure 1 illustrates the end-to-end architecture of ABPC-Net. The model processes input text through four sequential stages: (1) contextual embedding *via* frozen AraBERT, (2) sequential re-encoding *via* BiLSTM, (3) parallel multi-scale feature extraction via three independent CNN-Capsule branches with kernel sizes of 2,3 and 4 and (4) feature fusion and classification *via* concatenation and margin loss.

3.3.1 Transformer-based Embedding Layer

Our model builds upon the pre-trained Arabic transformer aubmindlab/bert-base-arabertv2 [21], which is a well-performing BERT model trained on an extensive dataset of Arabic text. Pre-processed text is tokenized with AutoTokenizer to obtain input_ids and attention_mask. The AraBERT encoder is kept fully frozen throughout all training stages (trainable = False), serving as a static contextual feature extractor. This design choice is justified on two complementary grounds. First, freezing the transformer prevents catastrophic forgetting of the broad linguistic knowledge encoded during pre-training on $\sim$ 77 GB of Arabic text, which would otherwise be at risk when fine-tuned on the comparatively smaller SANAD corpus ($\sim$ 100 K samples) [21]. Second, a frozen encoder provides a controlled experimental setting in which the contribution of the proposed downstream architecture, BiLSTM, parallel CNNs and Capsule projection, can be evaluated independently of the transformer representations, yielding a cleaner and more interpretable ablation framework. The empirical validation of this design choice is presented in Section 4.

3.3.2 Bidirectional LSTM Layer

After AraBERT generates contextualized embeddings, these sequences are fed into a recurrent neural network architecture that aims to capture longer-range dependencies as well as sequential constructs. Specifically, the full per-token output sequence of AraBERT, of shape (batch_size $\times 256 \times 768$), corresponding to the last hidden states of all 256 input tokens, is passed directly into the BiLSTM without any prior pooling, CLS token extraction or dimensionality reduction. This design preserves the complete positional and contextual information across all token positions, enabling the BiLSTM to model sequential dependencies that would otherwise be lost under aggregation. Long Short-Term Memory (LSTM) units are a kind of Recurrent Neural Network (RNN) set specifically for solving the vanishing gradient of typical RNNs. The LSTM units can learn a certain internal state at the cell level and through different gates (input, forget and output) is able to select to retain or forget content for a long amount of time. Because of their ability to store and retrieve information for arbitrary durations through gating mechanisms, LSTMs are well-suited for tasks involving sequential data, especially when temporal dynamics need to be modeled. We employ a BiLSTM in our model. This architecture enhances the standard LSTM by processing the input sequence in both forward and backward directions [22]. The outputs in both directions are combined, so that each word is more enriched than the words above it in light of both the preceding and following context.

3.3.3 Parallel Convolutional and Capsule Layers

The core innovation of our model lies in the parallel feature-extraction component, which is designed to effectively capture the different levels of textual features present in the input data. The sequence of hidden states generated by the BiLSTM layer is fed in parallel to the three parallel convolutional blocks for the feature-extraction process. Each convolutional block operates independently: the BiLSTM output is fed in parallel to three Conv1D layers (kernel sizes 2, 3 and 4), each followed by its own Global Max

Pooling layer and a dedicated Capsule projection layer. The three resulting capsule tensors - each of shape $(N \times D)$ where N is the number of target classes and $D = 16$ is the capsule dimension - are subsequently concatenated along the last axis, yielding a combined representation of shape ($N \times 3D$). This design ensures that bigram, trigram and quadrigram features contribute distinct vector representations to the final classification, rather than being merged prior to capsule projection. Each block consists of the following components:

- 1D Convolutional Layer (Conv1D): A Conv1D layer with kernel sizes of 2, 3 and 4 for the three respective parallel blocks [23] operates as an n-gram detector. It slides a filter over the sequence to identify various local patterns and features, such as adjacent word pairs (bigrams), trigrams and quadrigrams. The use of different kernel sizes allows for the extraction of features at multiple granularities.
- Global Max Pooling Layer: Following the Conv1D operation, the Global Max Pooling operation is applied to the output of the Conv1D layer. This layer aggregates the strongest feature present in the entire sequence, obtained by any filter used in the Conv1D operation, to capture the strongest local indicator for the particular feature. This layer represents a dimensionality-reduction operation, capturing the strongest n-gram features, which are then represented by the capsule layers in the form of vector representations for modeling the class-specific properties.
- Lightweight Vector-based Projection Head (Capsule-inspired): To increase representational expressiveness beyond standard scalar classification heads, we introduce a lightweight vector-based projection head inspired by capsule network principles [24]. We wish to be precise: this component is not a canonical capsule network, it involves no dynamic routing, no agreement mechanism and no part-whole relationship modeling in the sense of Sabour et al. [24]. Rather, it performs a deterministic learned linear projection of pooled CNN features into structured $N \times D$ vector representations, followed by a squashing non-linearity that bounds vector magnitudes while preserving directional information. The key distinction from a standard Dense classification head is that this projection produces multi-dimensional class-specific vectors rather than scalar logits, enabling vector-length-based class activation that preserves richer feature structure across the three parallel CNN branches.

Formally, given an input feature vector $\mathbf{x} \in \mathbb{R}^d$, the capsule layer performs a learned linear projection defined as:

$$\mathbf{u}_{\text{flat}} = \mathbf{x}\mathbf{W} \tag{1}$$

where $\mathbf{W} \in \mathbb{R}^{d \times (N \cdot D)}$ denotes the trainable weight matrix, N represents the number of capsules corresponding to the number of target classes and D defines the dimensionality of each capsule. The resulting projection $\mathbf{u}_{\text{flat}}$ is subsequently reshaped into structured capsule vectors:

$$\mathbf{U} = \text{reshape}(\mathbf{u}_{\text{flat}}, (N, D)) \tag{2}$$

To preserve vector magnitudes while maintaining directional information, a squashing nonlinearity is applied such that capsule lengths are bounded and interpretable:

$$\mathbf{V}_i = \frac{\|\mathbf{U}_i\|^2}{1 + \|\mathbf{U}_i\|^2} \frac{\mathbf{U}_i}{\sqrt{\|\mathbf{U}_i\|^2 + \epsilon}} \tag{3}$$

where $\mathbf{U}_i$ and $\mathbf{V}_i$ denote the i-th capsule before and after squashing, respectively. This non-linearity allows vector lengths to encode the strength of class activation while preserving multi-dimensional semantic information in vector orientations. Compared to conventional softmax classifiers that compress representations into scalar probability values, capsule vectors retain richer feature structures by maintaining multi-dimensional embeddings. This property enables the model to preserve semantic associations discovered during feature extraction, particularly when integrating outputs obtained from multi-scale convolutional branches. Furthermore, the deterministic projection mechanism eliminates the instability and computational overhead associated with routing iterations, resulting in more stable gradient behavior and reduced training cost. Final class predictions are derived from capsule vector lengths, $y_i = \|\mathbf{V}_i\|$, which aligns with capsule network principles and provides an interpretable vector-length-based decision mechanism. When BiLSTM precedes the CNN-Capsule pipeline, the hidden states, it produces encode directional and contextual dependencies across all token positions; even after

CNN and Global Max Pooling, the resulting vector retains structured inter-feature relationships that the capsule projection can meaningfully encode into vector-length-based class activations - a representation that is particularly suited to Arabic-text classification, where overlapping feature distributions across semantically similar categories (e.g., Politics *vs.* Finance, Tech *vs.* Finance) benefit from multi-dimensional encoding rather than scalar logit compression. In contrast, without BiLSTM, Global Max Pooling reduces the CNN output to a bag of independently selected scalar activations, discarding all sequential and relational structures. The capsule layer therefore functions as a relational feature-fusion mechanism the effectiveness of which is conditioned on the structural richness of its input.

The decision to maintain strict branch independence prior to capsule projection is grounded in three principled considerations. First, each kernel size (2, 3, 4) is designed to capture a structurally distinct linguistic granularity: bigrams can capture short morphological clitic combinations, trigrams can capture stem-affix patterns and short collocations and quadrigrams can capture longer phrasal units. Allowing cross-branch interaction prior to capsule projection would force the network to learn a single shared representation across these granularities, weakening the inductive bias that motivates the multi-scale design. Second, the capsule projection encodes directional information among feature dimensions through the learned weight matrix W; training this matrix on a structurally homogeneous within-branch input yields coherent within-granularity directional patterns, whereas mixing kernel sizes prior to projection forces the matrix to encode incompatible directional patterns simultaneously, weakening the relational signal that vector-based capsule representations are designed to preserve. Third, strict branch independence prior to fusion is the standard design pattern in multi-channel CNN architectures for text classification originating with Kim [23] and adopted in subsequent multi-scale text classifiers; our design extends this convention by inserting an independent capsule projection per branch before concatenation. The capsule-projection procedure is summarized in Algorithm 1.

Algorithm 1: Capsule-inspired Vector Projection

Input: Feature vector $x \in \mathbb{R}^d$ obtained from Global Max Pooling

Data: Learned projection matrix $\mathbf{W} \in \mathbb{R}^{\boldsymbol{d}\times(\boldsymbol{N}\times\boldsymbol{D})}$

Output: Squashed capsule matrix $\boldsymbol{V} \in \mathbb{R}^{N\times D}$

// Step 1: Linear projection to capsule space

$\mathbf{u}_{flat} = \mathrm{x}\mathbf{W}$

// Step 2: Structural reshaping into *N* class capsules

$\mathbf{U} = \mathrm{reshape}(\mathbf{u}_{flat}, (N, D))$

// Step 3: Non-linear capsule squashing

for $i \leftarrow 1\ to\ N$ **do**

$s_i = \|\mathrm{U}_{\boldsymbol{i}}\|^{\mathbf{2}}$ // Squared norm of the i[th] capsule

$\mathrm{V}_i = \frac{s_i}{1+s_i} \cdot \frac{\mathrm{U}_i}{\sqrt{s_i+\epsilon}}$ // Stable squashing function

end

return V

3.4 Classification Layer

Following the parallel feature-extraction pipeline, the three branch-specific capsule tensors, each of shape $(N \times D)$, where N is the number of target classes and $D = 16$ is the capsule dimension, are concatenated along the capsule-dimension axis to form a unified representation of shape ($N \times 3D$). To prevent overfitting, this concatenated tensor is passed through a Dropout layer with a rate of 0.3 applied along the feature axis. The actual fusion across kernel sizes is performed by a Lambda layer that

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

computes the Euclidean L_2-norm along the capsule-dimension axis, reducing the ($N \times 3D$) tensor to an N-dimensional class-activation vector. This norm operation aggregates the squared contributions of all $3D$ dimensions per class, treating the bigram, trigram and quadrigram capsule sub-vectors as additive evidence sources that jointly determine each class's activation magnitude. The resulting class-activation vector is supplied directly to the margin-loss function (with $m^+ = 0.9, m^- = 0.1$ and $= 0.5$), which trains the network, so that the correct-class capsule norm exceeds m^+while incorrect-class norms remain below m^-. At inference time, the predicted class is determined by arg max over the N capsule norms. The Euclidean norm provides a parameter-free, magnitude-preserving fusion that respects the vector-based semantics of capsule representations and aligns directly with the margin-loss framework.

3.5 Experimental and Hyper-parameter Settings

We performed all the experiments utilizing the TensorFlow and Keras DL frameworks in a Google Colab environment powered by an NVIDIA GPU. For the text analysis, we used the pre-trained Bert-Basearabertv 2 model. To allow efficient computation and use pre-learned linguistic features, we froze the BERT layers and used them as a static feature extractor. The input sequences were tokenized and padded to a maximum length of 256 tokens. The architecture of the $ABPC$-Net model consists of a Bidirectional LSTM (BiLSTM) layer with 128 units, followed by three parallel 1D-Convolutional layers. These layers utilize kernel sizes of 2, 3 and 4, respectively, with each employing 64 filters to effectively capture multi-scale n-gram features. A series of parallel Capsule layers was configured, each with a vector dimension of 16, ensuring consistent representational capacity across all feature-extraction branches. For model optimization, the Adam optimizer was employed with a learning rate of 0.001.

To improve class separability, we used a customized margin-loss function (with $m^+ = 0.9, m^- = 0.1$, $\lambda = 0.5$). Specifically, this loss function penalizes class capsules with low magnitudes for the correct class or large magnitudes for incorrect classes, pushing the model to learn more discriminative and distinct feature boundaries. The training process was limited to a maximum of 10 epochs with a batch size of 32. To prevent overfitting, an Early Stopping mechanism was implemented with a patience of 3 epochs, monitoring validation accuracy. Detailed hyper-parameter values for the experimental evaluation are summarized in Table 2.

Table 2. Hyper-parameters used in the ABPC-Net model.

Hyper-parameter	Value
Transformer	AraBERT v2
AraBERT Encoder	Trainable = False
Max Sequence Length	256
Batch Size	32
Epochs	10
Optimizer	Adam
Learning Rate	0.001
BiLSTM Units	128
CNN Filters	64
CNN Kernel Sizes	2, 3, 4
Capsule Dimension	16
Loss Function	Margin Loss

To ensure statistical reliability, all experiments involving ABPC-Net and the fine-tuned baseline models were repeated over five independent runs with different random seeds. Results are reported as mean accuracy ± standard deviation. The five random seeds used for the independent runs are 42, 123, 456, 789 and 2024. From a computational perspective, ABPC-Net comprises approximately 136 M total parameters, of which only ~ 1.09M are trainable, while the remaining ~ 135M correspond to the frozen AraBERT encoder. This design substantially reduces the trainable parameter budget relative to full fine-tuning approaches. The frozen AraBERT encoder occupies approximately 540 MB of GPU memory,

which constrains achievable batch sizes on consumer GPUs. Inference under GPU batched conditions averages 1.97 ms per sample, which is suitable for asynchronous applications, such as news categorization, content moderation and information retrieval. Under CPU inference, however, latency rises substantially, which may limit applicability for synchronous real-time pipelines requiring sub-10 ms response. Scaling ABPC-Net to larger workloads or stricter latency budgets would benefit from model distillation, quantization or substitution of the BiLSTM with a lighter temporal convolutional network.

4. RESULTS AND DISCUSSION

The ABPC-Net model was trained and evaluated using Akhbarona, AlArabiya, AlKhaleej and SANAD datasets. Model performance was evaluated using Accuracy, Precision, Recall and F1-score [25]. Figure 2 illustrates the training and validation behavior of ABPC-Net across all datasets. A consistent gap between training and validation accuracy is observed, particularly on SANAD and Akhbarona.

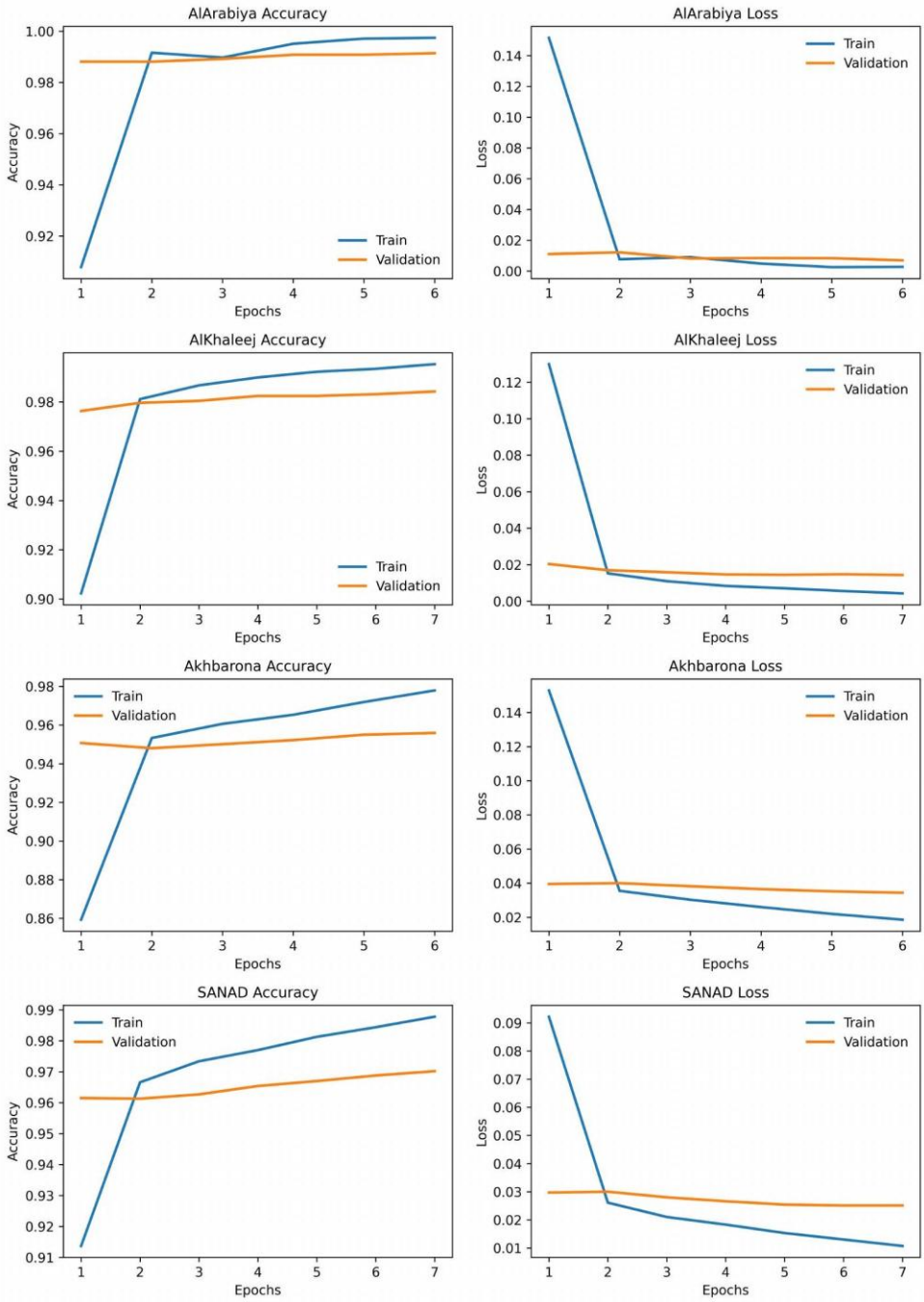


Figure 2. Training and validation performance across AlArabiya, AlKhaleej, Akhbarona and SANAD datasets. Note: Curves correspond to a single run of ABPC-Net.

This behavior is primarily attributable to two architectural factors rather than systematic overfitting: the Dropout layer (rate $= 0.3$) is active during training, but disabled during validation and the margin-loss function enforces stricter class boundary constraints during training (requiring capsule activations ≥ 0.9 for correct classes and ≤ 0.1 for incorrect classes) than are reflected in standard validation accuracy. Validation accuracy remained stable across epochs and the Early Stopping mechanism (patience $= 3$) restored the best-performing weights. The five-run statistical analysis further confirms the stability of validation performance ($\leq 0.12\%$ across all datasets). We compared ABPC-Net against four baseline configurations: frozen MARBERT, frozen AraBERT, MARBERT Full FT and AraBERT Full FT, each using a Dense classification head. All experiments involving ABPC-Net and the fine-tuned baselines were repeated over five independent runs using identical random seeds and experimental settings. Results are reported as mean ± standard deviation and paired-sample t -tests on the per-run accuracies confirmed that ABPC-Net's improvements are statistically significant ($p < 0.001$) on every evaluated dataset. Table 3 summarizes the comparative results under identical pre-processing, dataset splits and hardware conditions for directly re-implemented baselines (frozen MARBERT, frozen AraBERT, MARBERT Full FT, AraBERT Full FT).

Table 3. A comparative analysis of accuracy between ABPC-Net model and baseline models.

Methods	AlArabiya	AlKhaleej	Akhbarona	SANAD
MARBERT (frozen)	95.63 ± 0.13	95.78 ± 0.09	91.15 ± 0.11	90.60 ± 0.05
AraBERT (frozen)	96.79 ± 0.09	95.94 ± 0.10	91.74 ± 0.12	94.20 ± 0.06
MARBERT Full FT	95.98 ± 0.24	94.60 ± 0.10	90.23 ± 0.11	92.19 ± 0.04
AraBERT Full FT	97.52 ± 0.06	96.72 ± 0.10	93.49 ± 0.13	94.71 ± 0.05
ABPC-Net (ours)	$\mathbf{99.14 \pm 0.10}$	$\mathbf{98.40 \pm 0.10}$	$\mathbf{95.59} \pm 0.12$	$\mathbf{97.00 \pm 0.04}$

*All values are reported in percentages (%).
**The bold values indicate the high-accuracy performance achieved in each comparison.

On the aggregated SANAD dataset, ABPC-Net achieves $97.00 \pm 0.04\%$, surpassing the strongest fine-tuned baseline, AraBERT Full FT ($94.71 \pm 0.05\%$), by a margin of 2.29% and outperforming frozen AraBERT ($94.20 \pm 0.06\%$) by 2.80%. The consistently low standard deviation of ABPC-Net ($\leq$ 0.04% on SANAD) underscores the reproducibility of the proposed architecture. A similar pattern is observed across all individual sub-sets: ABPC-Net achieves $99.14 \pm 0.10\%$ on AlArabiya, $98.40 \pm 0.10\%$ on AlKhaleej and $95.59 \pm 0.12\%$ on Akhbarona, outperforming all baseline configurations in each case.

A particularly noteworthy finding is that ABPC-Net with a frozen AraBERT encoder consistently outperforms AraBERT Full Fine-tuned, which updates all 135 M transformer parameters with a simple Dense classification head, by margins of $2.29\%, 1.62\%, 2.10\%$ and 1.68% on SANAD, AlArabiya, Akhbarona and AlKhaleej, respectively. This result indicates that the performance gains of ABPC-Net are attributable to its structured downstream architecture, comprising BiLSTM sequential re-encoding, parallel multi-scale CNNs and a Capsule-inspired vector projection head, rather than to transformer fine-tuning. This empirical evidence further supports the frozen encoder design choice introduced in Sub-section 3.3.1, indicating that the structured downstream architecture compensates for what full fine-tuning achieves with a simple classification head. Furthermore, MARBERT Full FT ($92.19 \pm 0.04\%$ on SANAD) underperforms even frozen AraBERT ($94.20 \pm 0.06\%$), confirming that full fine-tuning of a larger transformer does not guarantee superior performance when the classification head lacks structural depth.

The class-wise performance metrics are detailed in Table 4, corresponding to a representative run selected based on its proximity to the mean accuracy across five independent runs. High precision and recall values across all categories confirm that the model maintains strong per-class discriminability, with particularly robust performance on Sports and Religion categories ($\geq$ 98% F1-score across all datasets) and greater variability in semantically overlapping categories, such as Politics and Finance, as further analyzed in Sub-section 4.4.

Table 4. Classification metrics for SANAD, AlArabiya, AlKhaleej and Akhbarona datasets.

Dataset	Class	Precision	Recall	F1-score
AlArabiya	Finance	99.18	98.11	98.64
	Medicine	99.19	99.73	99.46
	Politics	100	99.73	99.86
	Sports	99.19	99.73	99.46
	Tech	98.64	98.38	98.51
AlKhaleej	Culture	98.13	96.77	97.44
	Finance	99.22	98.00	98.61
	Medicine	98.17	99.08	98.62
	Politics	98.62	99.08	98.85
	Religion	97.41	98.46	97.93
	Sports	99.69	99.69	99.69
	Tech	98.45	97.85	98.15
Akhbarona	Culture	94.29	96.12	95.20
	Finance	91.78	93.28	92.52
	Medicine	95.38	98.66	96.99
	Politics	93.91	89.70	91.76
	Religion	98.65	98.51	98.58
	Sports	99.55	98.21	98.87
	Tech	97.99	94.63	96.28
SANAD	Culture	95.82	95.61	95.71
	Finance	94.70	96.15	95.42
	Medicine	97.60	98.82	98.21
	Politics	97.79	94.14	95.93
	Religion	96.64	97.95	97.29
	Sports	99.52	99.05	99.29
	Tech	97.34	97.28	97.31

4.1 Ablation Analysis

An ablation study was performed by systematically removing some components to assess the efficacy of the proposed architecture. In the follow-up, we review and critically compare model setting and model performance data with the goal of separating the impact of BiLSTM, CNN and Capsule Network layers. Table 5 summarizes the ablation study considering the influence of different components of the model on performance over datasets. The models are:

- AraBERT (Baseline): This is also used as the baseline of our method. The system established a benchmark accuracy of 94.19% on the aggregated dataset; it received scores of 91.75%, 96.81% and 95.91% on the Akhbarona, AlArabiya and AlKhaleej sub-sets, respectively. All subsequent models are evaluated against this performance.
- AraBERT + BiLSTM: The addition of a BiLSTM layer yielded significant improvements across the board. On the SANAD, the accuracy increased by 1.45% to 95.64%. The most substantial impact was observed on the Akhbarona sub-set, where performance climbed by 2.88% from 91.75% to

94.63%. This confirms the value of modeling sequential context, especially for more complex datasets.

- AraBERT + CNN: Augmenting the baseline with CNN layers resulted in the most significant performance gain from a single component. It increased the SANAD score by 2.14% to 96.33%. Notably, its performance on AlArabiya (98.70%) and AlKhaleej (98.00%) demonstrates the strength of CNNs in extracting highly informative local features.
- AraBERT + CNN + Capsule: The model that incorporates the AraBERT model with CNN and then adds the Capsule layer achieved an accuracy of 95.93% on the aggregated dataset. What is interesting to note is that this model achieved an accuracy that was 0.40 percentage points lower than the AraBERT-CNN model. This result reveals an important architectural insight: the capsule-projection layer is a relational feature-fusion mechanism the effectiveness of which is conditioned on the structural richness of its input. Without BiLSTM, Global Max Pooling reduces the CNN output to a bag of independently selected scalar activations, discarding all sequential and relational structures; projecting such a vector through a capsule layer cannot recover this discarded information and the additional parameters introduce optimization noise that marginally degrades performance. In contrast, when BiLSTM is present in the full ABPC-Net pipeline, the capsule layer receives structurally richer input and its contribution becomes clearly positive: ABPC-Net (97.02%) outperforms AraBERT + BiLSTM + CNN (95.89%) by 1.13% on SANAD, confirming that the capsule projection effectively encodes relational feature structure when provided with sufficiently rich sequential input.
- AraBERT + BiLSTM + CNN: This model was a direct integration of components in our analysis. It reported 95.89% accuracy on the SANAD dataset. This, importantly, is 0.44% less than that achieved by the AraBERT + CNN model. The same performance degradation was observed in other sets, such as AlArabiya (0.75% drop) and AlKhaleej (0.42% drop). This is the quantitative evidence that the naive combination of these layers is non-optimal and doesn't properly harmonize the features being extracted.
- ABPC-Net: Finally, in this part, our proposed architecture alleviated degradation observed in the previous two architectures. In the case of the SANAD dataset, an accuracy of 97.02% was obtained by a cleverly-integrated parallel feature output from the BiLSTM and the CNN layers *via* the Capsule Network. It indicates 0.69% improvement over the best single-component model and 1.09% over the AraBERT + CNN + Capsule model. This model displayed the best performance in all the individual sub-sets.

Table 5. Ablation analysis on AlArabiya, AlKhaleej, Akhbarona and SANAD datasets.

Methods	AlArabiya	AlKhaleej	Akhbarona	SANAD
AraBERT	96.81	95.91	91.75	94.19
AraBERT + BiLSTM	98.00	96.88	94.63	95.64
AraBERT + CNN	98.70	98.00	95.14	96.33
AraBERT + CNN + Capsule	98.32	97.41	95.01	95.93
AraBERT + BiLSTM + CNN	97.95	97.58	94.88	95.89
ABPC-Net	99.14	98.42	95.59	97.02

In summary, this numerical analysis confirms that while BiLSTM and CNN operate effectively as feature-extraction components, their successful integration depends on the architectural context in which they are combined. In particular, the capsule-inspired projection is not simply an additional component, but acts as a fusion mechanism that enables structured integration of parallel feature streams. The comparison between AraBERT + BiLSTM + CNN (95.89%) and ABPC-Net (97.02%) provides an estimate of the effect of introducing a vector-based projection head between feature aggregation and classification. The observed improvement of approximately 1.13% on SANAD, along with consistent gains across sub-sets, indicates a representational benefit over scalar classification in this specific architectural setting. Furthermore, the vector-length-based decision mechanism provides an

interpretable class-activation signal, which may be beneficial in scenarios where feature boundaries are ambiguous. This result highlights that the effectiveness of the capsule-inspired projection is conditional rather than universal. Specifically, its performance gain emerges only when the input representation preserves sequential structure, as provided by the BiLSTM layer. This finding suggests that capsule-inspired mechanisms are more appropriately viewed as relational fusion operators than as standalone classifiers, particularly in Arabic-text classification settings.

4.2 Comparison with Recent Published Methods

The performance advantages of ABPC-Net over recent transformer-based models are attributable not to the transformer backbone itself, which is identical to the AraBERT baseline, but to the structured downstream architecture that better exploits the high-dimensional representations produced by the frozen encoder. To contextualize the performance of ABPC-Net against published Arabic-text classification methods, we present a comparison with a set of recent models on SANAD, AlArabiya, AlKhaleej and Akhbarona datasets in Table 6. We emphasize at the outset that the results in Table 6 for competing methods are reproduced directly from their respective original publications and were not re-implemented by the authors under identical experimental conditions. Although the standard SANAD train/test split is widely adopted across these works, the external results may still differ in pre-processing pipelines, validation protocols, hyper-parameter tuning, framework implementations, hardware environments and in single-run *versus* multi-run reporting. On AlKhaleej sub-set, ABPC-Net reaches an accuracy of $98.40 \pm 0.10\%$, which compares favorably to the published results of CNN with character-level model [14] (98.00%) and Tasneef [11] (97.49%). This result reflects the effectiveness of the hybrid architecture in classifying articles from this news source, with consistent gains over CNN and character-level feature-extraction approaches. Furthermore, a recently proposed model integrating Graph Convolutional Networks (GCNs) with AraBERT embeddings [26] achieves 97.25% on AlKhaleej, representing a qualitatively different architectural direction from sequence-based methods. ABPC-Net shows a margin of 1.15% relative to this graph-based model (98.40% *vs.* 97.25%).

Table 6. Accuracy comparison of the ABPC-Net model against SOTA methods.

Methods	AlArabiya	AlKhaleej	Akhbarona	SANAD
BiGRU [5]	97.41	96.46	92.23	94.83
Attention-GRU [5]	96	96.66	92.95	94.98
CGRU [5]	97.19	96.86	94	95.71
ArCAR [10]	-	97.47	-	-
CNN with character level [16]	-	98	-	-
Transformer-CNN [27]	97.19	96.55	92.14	94.29
Tasneef [11]	98.43	97.49	95.43	-
TCAODL-ANA [28]	-	-	-	95.48
Inception-CNN + LSTM [4]	82	96	92	92
GCN+AraBERT [26]	-	97.25	-	-
ABPC-Net	$\mathbf{99.14 \pm 0.10}$	$\mathbf{98.40 \pm 0.10}$	$\mathbf{95.59 \pm 0.12}$	$\mathbf{97.00 \pm 0.04}$

*All values are reported in percentages (%).
**The bold values indicate the high-accuracy performance achieved in each comparison.
*** Results for external models are sourced from original publications and may reflect different pre-processing pipelines, train/test splits or hardware environments.

For AlArabiya dataset, our model achieves strong performance in classification with an accuracy of 99.14%. That's 0.71% better than Tasneef's [11] previous best result (98.43%). Thus, this indicates a level of accuracy; it demonstrates that the ABPC-Net method has learned the linguistic structures of AlArabiya dataset, indicating that contextual-local-hierarchical feature combinations have proven effective. On the more challenging Akhbarona sub-set, ABPC-Net achieves an accuracy of $95.59 \pm 0.12\%$, which compares favorably to the published results of Tasneef [11] (95.43%), CGRU [5]

(94.00%) and Transformer-CNN [27] (92.14%). The consistency of these numerical advantages across multiple comparison points suggests that the architecture handles Akhbarona's linguistic variability effectively. Finally, on the aggregated SANAD corpus, ABPC-Net achieves an accuracy of $97.00 \pm 0.04\%$, which compares favorably to the published results of CGRU [5] (95.71%) and TCAODL-ANA [28] (95.48%), showing a margin of approximately 1.31% relative to the closest competing entry in this comparison. Taken together, these results position ABPC-Net as a strong-performing approach for Arabic news classification on the SANAD benchmark, though generalization to other datasets and domains warrants further investigation.

4.3 Intra-domain Transferability and Few-shot Adaptation

To assess the intra-domain transferability of ABPC-Net beyond the SANAD benchmark, we conducted a two-phase cross-dataset evaluation using BBC Arabic and CNN Arabic [29], two widely-used Arabic news corpora that, while distinct from SANAD in source, writing style and category structure, remain within the news domain. Accordingly, the experiments in this section evaluate robustness to source-level distribution shift.

Phase 1: Zero-shot Cross-dataset Evaluation

The ABPC-Net model trained exclusively on SANAD was evaluated directly on BBC Arabic (7 categories) and CNN Arabic (6 categories) without any additional training. Since both datasets do not contain the Medical and Religion categories present in SANAD, a systematic category mapping was applied to align label spaces across datasets. Specifically, BBC Arabic categories were mapped as follows: اقتصاد و اعمال→ Finance, العالم → and Politics. CNN Arabic categories were mapped as follows: business → Finance, علوم وتكنولوجيا → Tech, رياضة → Sports, منوعات and عرض الصحف→ Culture, اخبار العالم and اخبار الشرق الاوسط → Politics. It is worth noting that the merging of world and middle_east into a single Politics class, as well as منوعات and عرض الصحف into Culture, introduces label-space asymmetry that partially accounts for the performance variation observed across categories in the zero-shot evaluation. The category mapping applied above introduces three sources of bias that should be considered when interpreting the zero-shot results. First, merging multiple BBC and CNN categories into single SANAD classes creates merged classes that are broader than their SANAD counterparts, which may affect per-class metrics. Second, the merged classes are semantically broader than the original SANAD definitions, which may shift accuracy in either direction depending on alignment with the model's learned class boundaries. Third, both BBC Arabic and CNN Arabic lack the Medical and Religion categories present in SANAD, producing labelspace asymmetry that the zero-shot evaluation cannot fully resolve. Consequently, the reported zero-shot accuracies should be interpreted as they are partially confounded by these label-mapping effects.

Under zero-shot conditions, ABPC-Net achieved 60.76% on BBC Arabic and 75.33% on CNN Arabic. These results reflect the combined effect of source-level distribution shift between news outlets and the label-mapping bias described above. The substantial gap relative to in-domain SANAD performance should therefore be attributed to both factors and the absolute zero-shot values should not be treated as direct measures of cross-source generalization. Importantly, categories with universal linguistic patterns transferred well, Sports achieved 96.99% F1 on CNN Arabic and Politics achieved 83.50% F1, while domain-specific categories, such as Tech (39.31% F1)و showed greater sensitivity to cross-source variation, a finding consistent with cross-domain transfer literature in Arabic NLP. The observed performance gap is further attributable to the absence of Medical and Religion categories in the target datasets, which introduces systematic label-space mismatch.

Phase 2: Few-shot Domain Adaptation

To evaluate adaptability under low-resource conditions, ABPC-Net was fine-tuned using only 20% of BBC Arabic and CNN Arabic, respectively, with the remaining 80% reserved for testing. This protocol simulates a realistic deployment scenario where limited target-domain supervision is available. Following adaptation with only 20% of target domain data, ABPC-Net achieved 94.36% on BBC Arabic and 89.20% on CNN Arabic, demonstrating substantial performance recovery with minimal additional supervision. These results confirm that the ABPC-Net architecture, with its frozen AraBERT encoder and trainable downstream layers, is particularly well-suited for rapid domain adaptation, as only the lightweight BiLSTM-CNN-Capsule pipeline requires updating. Notably, the few-shot adaptation

protocol yielded a performance gain of 33.60 percentage points on BBC Arabic (from 60.76% to 94.36%) and 13.87 percentage points on CNN Arabic (from 75.33% to 89.20%), demonstrating that ABPC-Net's frozen AraBERT encoder retains transferable linguistic representations while the trainable downstream layers, BiLSTM, CNN and Capsule, adapt rapidly to the target domain distribution with minimal supervision. Taken together, these results indicate that ABPC-Net's generalization is bounded, but adaptable: zero-shot transfer is constrained by source-level distribution shift and label-space asymmetry, while few-shot finetuning efficiently bridges these gaps, recovering the majority of source-specific performance using only 20% of target-domain data.

Table 7. Cross-dataset generalization results on BBC Arabic and CNN Arabic.

Evaluation Protocol	Dataset	Accuracy	Training Data
Zero-shot Transfer	BBC Arabic	60.76	SANAD only
Zero-shot Transfer	CNN Arabic	75.33	SANAD only
Few-shot Adaptation	BBC Arabic	94.36	SANAD + 20% BBC
Few-shot Adaptation	CNN Arabic	89.20	SANAD + 20% CNN
Independent Training	BBC Arabic	99.37	BBC only
Independent Training	CNN Arabic	94.68	CNN only

*All values are reported in percentages (%).
** Independent Training results use 70%/10%/20% train/validation/test split.
*** Medical and Religion categories absent in BBC and CNN datasets.

The architecture therefore exhibits consistent intra-domain transferability across Arabic news sources and rapid few-shot adaptability under low-resource conditions. We emphasize, however, that these findings do not establish cross-domain generalization beyond the news genre; evaluation on dialectal Arabic, conversational text, scientific Arabic and other non-news domains remains an important direction for future work and is identified explicitly in the Conclusion. Table 7 summarizes the cross-dataset generalization results.

4.4 Error Analysis

The overall result of the text-classification model performance was highly accurate and the misclassifications were due to some semantic overlap of certain categories, as seen from the text-training dataset. This work is based on the confusion matrix illustrated by Figure 3, which demonstrates the proportion of correct and incorrect predictions. For AlArabiya dataset, as shown in Figure 3(a), a clustering of mistakes appears in the Tech category, which had 2 misclassifications as Finance, 2 as Medical and 2 as Sports, according to the confusion matrix. Similarly, the Finance category had a misclassification of the same case as Tech, only 1. While indicating high overall accuracy, this highlights that there is a slight overlap between the categories with similar terminology.

In AlKhaleej dataset, as shown in Figure 3(b), the confusion matrix reveals a specific focus on errors between the Finance and Tech categories. Specifically, 8 Finance articles were misclassified as Tech, while 4 Tech articles were classified as Finance. A clear example of this ambiguity is found in an article stating:"
استحوذت ميديا كويست، الرائدة في مجال الاعلام في منطقة الشرق الأوسط وشمال إفريقيا، على ما قدره 30% من أسهم الموقع الإلكتروني Whiteme.net."

This article was misclassified as Tech instead of Finance. The likely reason for this error is the dominance of technical terms such as "المواقع الإلكترونية" (websites) and the specific domain name "Whiteme.net". The model appears to have prioritized these technical features over the broader financial context established by key terms, like "أسهم" (stocks) and "استحوذت" (acquired).

An example is an article titled "حل جديد شحن اجهزه الكترونيه طاقه شمسيه" (discussing innovative solar-powered charging solutions for electronic devices), which was predicted as Tech instead of Finance. The text emphasizes technical details, such as "طاقه شمسيه" (solar energy), "شاحن" (charger) and "بطاريه ليثيوم" (lithium battery), which likely dominated the convolutional filters and aligned closely with Tech

vocabulary. The model's reliance on these technology-centric n-grams overshadowed subtle financial undertones, like market applications, leading the capsule projection mechanism to favor Tech Class.

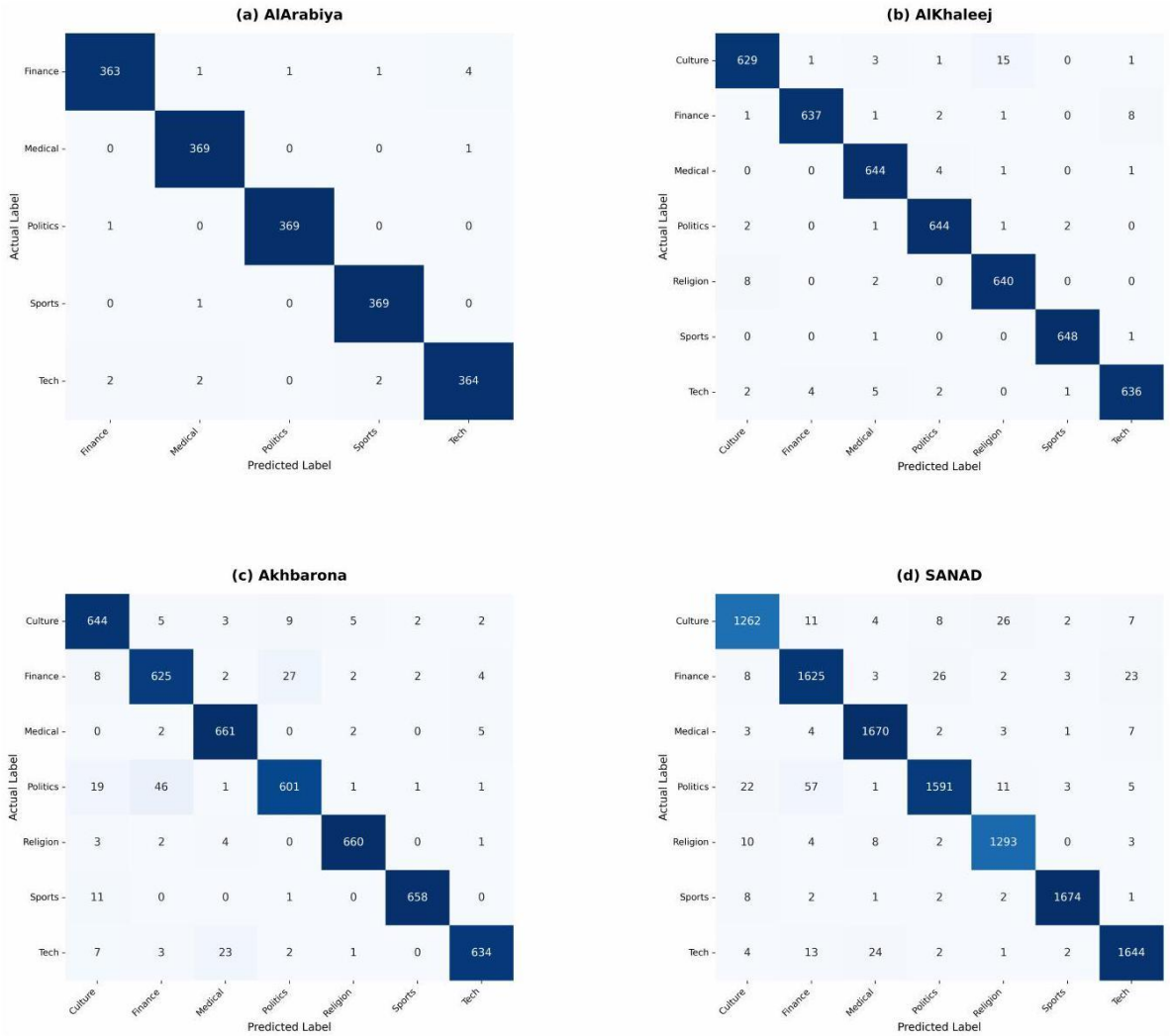


Figure 3. Confusion matrices for AlArabiya (a), AlKhaleej (b), Akhbarona (c) and SANAD (d) datasets, illustrating per-class classification performance and primary sources of misclassification.

As for Akhbarona dataset, Figure 3(c), the confusion matrix highlights prominent errors in the Politics category, with 46 cases misclassified as Finance and 19 as Culture, indicating strong semantic overlap between these domains. For example, an article titled: "ميزانية وزارة الاوقاف ارتفعت بأزيد من 2000 بالمائة", was misclassified as Finance instead of Politics. This error is attributed to the presence of strong financial terms, like "ميزانية" (Budget) in the title, along with "الاستثمار" (Investment) and "السنة المالية" (Fiscal year) in the body text, which outweighed the political context of the ministry's activities in the model's decision-making process.

Finally, for the aggregated SANAD dataset, Figure 3(d), the confusion matrix indicates that the combination of data increased the complexity of the data. Among the many examples, there is a notable focus on misclassification of Politics as Finance with 57 cases of Politics classified as Finance. This illustrates the complexity of reconciling differing contexts when combining diverse sources. Data integration and better generalization were achieved, but also exacerbated specific ambiguities, such as lingering Tech and Finance exchanges.

4.5 Limitations and Dataset Bias

Although ABPC-Net achieves strong empirical performance on the evaluated benchmarks, several important limitations should be acknowledged for accurate interpretation of the results. These limitations relate primarily to dataset bias and to the scope of the experimental evaluation.

First, all evaluated datasets, SANAD, BBC Arabic and CNN Arabic, consist exclusively of Modern Standard Arabic (MSA) news text. However, Arabic encompasses a wide spectrum of dialects and registers, including Egyptian, Levantine, Gulf and Maghrebi varieties, as well as informal written forms commonly used in social media and conversational platforms. These variants differ substantially from MSA in morphology, syntax and lexical usage. ABPC-Net has not been evaluated on such dialectal or informal data and its performance in these settings remains an open question.

Second, because ABPC-Net employs a frozen AraBERT encoder, it inherits the pre-training distribution of AraBERT, which is predominantly based on MSA corpora. As a result, the effectiveness of the proposed downstream architecture is closely related to this representation space. Extending the approach to dialectal Arabic would likely require the use of dialect-aware or multi-dialect pre-trained encoders.

Third, all evaluated datasets are drawn from the news domain, which tends to emphasize political, financial and sports content while under-representing other genres, such as scientific, technical, legal and conversational text. Consequently, the reported performance should be interpreted within the scope of Arabic news classification rather than as a general indicator of performance across all Arabic-text domains.

Fourth, SANAD dataset itself reflects specific editorial styles associated with Gulf and pan-Arab news portals (AlArabiya, AlKhaleej and Akhbarona), which differ from other regional styles. This contributes to the source-level distribution shift observed in the cross-dataset experiments reported in Sub-section 4.3.

Fifth, the observed drop in zero-shot performance between SANAD and external datasets (BBC Arabic and CNN Arabic) further highlights the sensitivity of the model to distribution shift, even within the news domain. While few-shot adaptation substantially improves performance under limited supervision, these results suggest that careful adaptation remains important when transferring to new sources.

Sixth, the ablation study compares the capsule-inspired projection against a scalar baseline (AraBERT+BiLSTM + CNN with direct margin loss) but does not exhaustively compare against alternative post-pooling mechanisms, such as deeper Dense heads, gating or branch-level attention. Whether similar gains could be achieved using lower-complexity alternatives remains an open question.

Overall, these limitations indicate that the reported results demonstrate strong performance within the scope of MSA Arabic news classification, while broader evaluation across dialects, domains and genres, as well as direct comparison against simpler architectural alternatives, remains an important direction for future work.

5. CONCLUSION

In this article, we have presented ABPC-Net as a structured downstream architectural design for transformer-based Arabic-text classification. The contribution is primarily architectural and empirical, supported by systematic experimental analysis. Their interaction yields non-trivial performance gains and provides practical insights into effective downstream design - particularly regarding the conditional effectiveness of capsule-inspired projections, which function as relational fusion operators rather than standalone classifiers. The capsule-inspired projection contributes consistent performance improvements when used within an appropriate architectural context, highlighting its role as a complementary component rather than a standalone solution.

We presented a rigorous ablation analysis showing that the fusion of these aspects is a key factor to performance. When combined, the ABPC-Net model demonstrates consistent improvements over individual baselines and hybrid configurations under the evaluated settings on the SANAD benchmark. Extensive evaluation yields a mean accuracy of $97.00 \pm 0.04\%$ on the full SANAD dataset, with particularly strong performance on AlArabiya ($99.14 \pm 0.10\%$), though these results should be interpreted within the scope of the tested datasets and experimental conditions. Furthermore, cross-dataset evaluation on BBC Arabic and CNN Arabic provides evidence of consistent intra-domain transferability across Arabic news sources and of rapid few-shot adaptability under low-resource conditions.

However, in spite of these encouraging findings, there were some limitations and drawbacks in our study. One such difficulty is the computational complexity induced by the hybrid structure, hierarchical

feature representation enabled by the capsule-inspired vector encoding mechanism, which delays the training of the model. On closer inspection, misclassification patterns persisted between closely related categories, such as Politics and Finance, indicating that contextual overlap remains challenging even for sophisticated architectures. Several limitations of the current work warrant acknowledgment. First, SANAD comprises exclusively Modern Standard Arabic news articles drawn from three specific portals, introducing source-specific stylistic and topical biases that may limit generalization to other domains or writing styles. The cross-dataset evaluation on BBC Arabic and CNN Arabic (Sub-section 4.3) provides empirical evidence of this constraint, where zero-shot transfer performance reflects the domain shift between the training distribution and unseen target sources. Second, as AraBERT is pre-trained predominantly on MSA corpora, ABPC-Net inherits an inherent limitation in handling dialectal Arabic variants, including Egyptian, Levantine and Gulf dialects, which differ substantially from MSA in morphology, vocabulary and syntactic structure. Extending the architecture to dialectal Arabic through dialect-aware pre-trained encoders or multi-dialect training data remains an important direction for future work.

In the future, we want to investigate the model compression based on knowledge distillation or quantization methods, thereby aiming to reduce computation demand and inference time without decreasing accuracy and making the model more practical in actual applications. In order to mitigate this issue of semantic overlap, we intend to explore more complex types of data augmentation, such as back-translation or contextual word replacement, to generate more diverse training samples to solve the problem of class pairs. Additionally, we extend the architecture to support dialectal Arabic variations (e.g., Egyptian, Levantine) and apply XAI methods (explainable AI) to visualize the routing decision of the Capsule Network, so that we might achieve greater transparency and wider impact on Arabic NLP tasks. A promising direction for future work is replacing the BiLSTM layer with a lighter temporal convolutional network (TCN) for sequence modeling, which may achieve similar representational capacity with lower computational cost. This would further explore the trade-off between sequential-modeling effectiveness and efficiency in transformer-based hybrid architectures.

REFERENCES

[1] A. Alrayzah, F. Alsolami and M. Saleh, "AraFastQA: A Transformer Model for Question-answering for Arabic Language Using Few-shot Learning," Computer Speech & Language, vol. 95, p. 101857, 2026.

[2] A. Wali et al., "Evaluating Arabic Sentiment Analysis with GPT-4o: A Comparative Study of Raw and Pre-processed Text," J. of Cases on Inf. Tech., vol. 28, no. 1, pp. 1-19, 2026.

[3] J. H. Yousif, "Artificial Intelligence and Machine Learning for Enhancing Arabic Fake News Detection: A BERT-based Transformer Approach," Procedia Computer Science, vol. 275, pp. 809-816, 2026.

[4] E. Alnagi, R. Ghnemat and Q. Abu Al-Haija, "Boosting Arabic Text Classification Using Hybrid Deep Learning Approach," Discover Applied Sciences, vol. 7, no. 6, p. 540, May 2025.

[5] A. Elnagar, R. Al-Debsi and O. Einea, "Arabic Text Classification Using Deep Learning Models," Information Processing & Management, vol. 57, no. 1, p. 102121, Jan. 2020.

[6] N. Boudad, R. Faizi, R. Oulad Haj Thami and R. Chiheb, "Sentiment Analysis in Arabic: A Review of the Literature," Ain Shams Engineering Journal, vol. 9, no. 4, pp. 2479-2490, Dec. 2018.

[7] B. M. Merzah, J. Razmara and Z. Salmanian, "Hybrid Deep Learning Models for Fake News Detection: Case Study on Arabic and English Languages," Frontiers in Big Data, vol. 8, DOI: 10.3389/fdata.2025.1683786, Jan. 2026.

[8] A. B. Nassif et al., "Arabic Fake News Detection Based on Deep Contextualized Embedding Models," Neural Computing & Applications, vol. 34, no. 18, pp. 16019-16032, Sep. 2022.

[9] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni and D. Nouvel, "Arabic Natural Language Processing: An Overview," J. of King Saud Uni.-Computer and Inf. Sciences, vol. 33, no. 5, pp. 497-507, 2021.

[10] A. Y. Muaad et al., "A Novel Deep Learning ArCAR System for Arabic Text Recognition with Character-level Representation," Computer Sciences and Mathematics Forum, vol. 2, no. 1, Article no. 14, DOI: 10.3390/IOCA2021-10903, and Presented at the 1st Int. Electronic Conf. on Algorithms, Sep. 2021.

[11] M. Louail et al., "Tasneef: A Fast and Effective Hybrid Representation Approach for Arabic Text Classification," IEEE Access, vol. 12, pp. 120804-120826, 2024.

[12] I. Jamaleddyn, R. ayachi and M. Biniz, "Novel Multichannel Deep Learning Model for Arabic News Classification," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 10, no. 4, pp. 453-468, DOI: 10.5455/jjcit.711720086134, Dec. 2024.

[13] Md. M. Hossain et al., "A Hybrid Attention-based Transformer Model for Arabic News Classification Using Text Embedding and Deep Learning," IEEE Access, vol. 12, pp. 198046-198066, 2024.

[14] M. El Kourdi, A. Bensaid and T. E. Rachidi, "Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm," Proc. of the Workshop on Computational Approaches to Arabic Script-based Languages (Semitic '04), pp. 51-58, Geneva, Switzerland, 2004.

[15] L. Al Qadi, H. El Rifai, S. Obaid and A. Elnagar, "A Scalable Shallow Learning Approach for Tagging Arabic News Articles," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 6, no. 3, pp. 263-280, DOI: 10.5455/jjcit.71-1585409230, 2020.

[16] A. Y. Muaad et al., "An Effective Approach for Arabic Document Classification Using Machine Learning," Global Transitions Proceedings, vol. 3, no. 1, pp. 267-271, Jun. 2022.

[17] D. Alsaleh and S. Larabi-Marie-Sainte, "Arabic Text Classification Using Convolutional Neural Network and Genetic Algorithms," IEEE Access, vol. 9, pp. 91670-91685, 2021.

[18] M. S. H. Ameur, R. Belkebir and A. Guessoum, "Robust Arabic Text Categorization by Combining Convolutional and Recurrent Neural Networks," ACM Trans. on Asian and Low-resource Language Information Processing, vol. 19, no. 5, pp. 1-16, DOI: 10.1145/3390092, Sep. 2020.

[19] A. A. Jalil and A. H. Aliwy, "Classification of Arabic Social Media Texts Based on a Deep Learning Multi Tasks Model," Al-Bahir, vol. 2, no. 2, DOI: 10.55810/2313-0083.1030, May 2023.

[20] B. B. Al-onazi et al., "Automated Arabic Text Classification Using Hyper-parameter Tuned Hybrid Deep Learning Model," Computers, Materials & Continua, vol. 74, no. 3, pp. 5447-5465, 2023.

[21] W. Antoun, F. Baly and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," Proc. of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pp. 9-15, Marseille, France, 2020.

[22] A. Jalili, H. Tabrizchi, J. Razmara and A. Mosavi, "BiLSTM for Resume Classification," Proc. of the 2024 IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 519-524, Stará Lesná, Slovakia, 2024.

[23] Y. Kim, "Convolutional Neural Networks for Sentence Classification," Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751, Doha, Qatar, DOI: 10.3115/v1/D14-1181, 2014.

[24] S. Sabour, N. Frosst and G. E. Hinton, "Dynamic Routing between Capsules," Proc. of the 31st Conf. on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., vol. 30, 2017.

[25] D. M. W. Powers, "Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation," arXiv: 2010.16061, 2020.

[26] M. Benhammouda, A. Khobzaoui and N. Mahammed, "Arabic Text Classification Using Graphs and Deep Learning," IJCESEN, vol. 11, no. 4, pp. 9415-9421, DOI: 10.22399/ijcesen. 4402, 2025.

[27] M. Berrimi, M. Oussalah, A. Moussaoui and M. Saidi, "A Comparative Study of Effective Approaches for Arabic Text Classification," Available at SSRN 4361591, DOI: 10.2139/ssrn. 4361591, Feb. 2023.

[28] M. S. A. Alzaidi et al., "Enhanced Automated Text Categorization *via* Aquila Optimizer with Deep Learning for Arabic News Articles," Ain Shams Engineering Journal, vol. 16, no. 1, p. 103189, DOI: 10.1016/j.asej.2024.103189, Jan. 2025.

[29] M. K. Saad and W. Ashour, "OSAC: Open Source Arabic Corpora," Proc. of the 6th ArchEng Int. Symposiums on Electrical and Electronics Engineering and Computer Science (EEECS'10), vol. 10, p. 55, DOI: 10.13140/2.1.4664.9288, 2010.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

ملخص البحث:

لا يزال تصنيف النّصوص العربية يمثل تحدياً مهماً بسبب الثّراء الصرفي والتّنوع اللّهجي والتّعقيد الدّلالي الذي تتميز به اللّغة العربية. وعلى الرّغم من التّقدم الذي حقّقته النّماذج المعتمدة على المحولات Transformers مثل AraBERT، فإنّ العديد من الأساليب ما تزال تعتمد على طبقات تصنيف بسيطة لا تستفيد بصورة كاملة من الخصائص الهرمية والتسلسلية للنّصوص.

تقدّم هذه الدّراسة إطاراً هجيناً بإسم ABPC-NET يجمع بين مشفر AraBERT مجمداً لاستخراج التّمثيلات السياقية، و شبكة ثنائية الاتجاه من نوع BiLSTM لنّمذجة الاعتماديات التسلسلية، وفروعاً متوازية من الشّبكات العصبية الالتفافية CNN متعدّدة المقاييس باحجام نوافذ (2 و3 و4) لاستخراج خصائص n-gram، بالإضافة إلى رأس إسقاط متّجهي مستوحى من شبكات الكبسولات Capsule Networks لدمج الخصائص الهرمية.

تمّ تقييم النّموذج على مجموعة بيانات SANAD ومجموعاتها الفرعية (العربية، والخليج، وأخبارنا) عبر خمس تجارب مستقلة. وحقّق متوسط دقّة بلغ 97.00±0.04% على SANAD، و 99.14±0.10% على العربية، و 98.40±0.10% على الخليج، و 95.59±0.12% على أخبارنا، متفوقاً بصورة متّسقة على نماذج AraBERT و MARBERT المجمدة والمضبوطة بالكامل ضمن الظروف التّجريبية نفسها.

كما أظهرت التّجارب العابرة لمجموعات البيانات على BBC Arabic و CNN Arabic قدرة النّموذج على الانتقال المعرفي داخل مجال الأخبار العربية والتّكيف السريع مع مصادر بيانات جديدة باستخدام عدد محدود من أمثلة التّدريب.

وتشير النّتائج إلى أنّ البنية المقترحة تمثل إطاراً معرفياً فعالاً وقابلاً للتّعميم لتّصنيف النّصوص الإخبارية العربية المكتوبة باللّغة العربية الفصحى الحديثة.

Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 12, No. 02, June 2026.

ANALYSIS OF PCAP-DERIVED FLOW-BASED TRAFFIC REPRESENTATION FOR LIGHTWEIGHT INTRUSION DETECTION

Andrés Eduardo Villamarín Olmos and Edward Paul Guillen Pinto

(Received: 13-Mar.-2026, Revised: 30-May-2026, Accepted: 6-Jun.-2026)

ABSTRACT

The proliferation of interconnected network infrastructures and IoT devices has significantly expanded the cyber-attack surface, requiring efficient Machine Learning-based Intrusion Detection Systems (IDSs). Although reference datasets like UNSW-NB15 exist, their official features impose limitations regarding flexibility and class imbalance. This study evaluates the impact of a custom data representation by constructing a new dataset from the original UNSW-NB15 PCAP files. We implemented a workflow to label packets, group unidirectional flows and extract a reduced set of 21 features, comparing this representation with the official 49-feature UNSW-NB15 set using different ML architectures in binary and multi-class classification tasks. Results indicate that the custom dataset achieves competitive performance despite a significant reduction in file size and the number of features. Notably, the custom representation effectively balances detection accuracy with computational efficiency, offering a viable strategy for environments with strict operational constraints, such as edge nodes or IoT gateways.

KEYWORDS

Intrusion detection systems (IDSs), Network traffic classification, UNSW-NB15, Machine learning, Network security.

1. INTRODUCTION

In the current digital era, the proliferation of connected devices has reached unprecedented levels. Industry reports project that the number of Internet of Things (IoT) devices will exceed 30 billion by 2030, generating a massive volume of data at the network edge [1]. This hyper-connectivity, however, has significantly expanded the attack surface, with global cybercrime costs estimated to reach $10.5 trillion annually by 2025 [2]. In this scenario, traditional signature-based security mechanisms are insufficient to handle the volume and sophistication of modern threats, making Intrusion Detection Systems (IDSs) based on Machine Learning (ML) indispensable tools for automating the identification of anomalous traffic.

However, the effectiveness of these ML-based systems largely depends not only on the algorithms used, but fundamentally on the quality of the data and the design of the features representing network traffic. While deep-learning models have shown high detection rates, their deployment in limited-resource environments, such as IoT gateways, is often hindered by the high computational cost associated with processing high-dimensional data. Consequently, optimizing the data representation stage is critical to achieving a balance between detection accuracy and operational efficiency.

Among contemporary datasets, UNSW-NB15 for enterprise network traffic has become a reference for the evaluation of detection methods, since it combines real traffic with multiple families of up-to-date attacks and provides CSV files with 49 attributes derived from PCAP captures. Nevertheless, the direct use of these official CSV files imposes certain constraints: on the one hand, it restricts the flexibility to implement custom packet grouping strategies; on the other, it inherits class imbalance and pre-processing decisions that shape the behavior of the models. Consequently, it remains an open question to what degree it is possible to redesign the representation of traffic without sacrificing performance, while at the same time reducing data size and the computational cost of training.

In this context, this study investigates how different choices in the representation and structuring of traffic data affect IDS performance. Using UNSW-NB15 as a case study, we construct an alternative packet-grouping strategy, reduce feature sub-sets and compare them against the original configuration in order to characterize their impact on detection performance and their suitability under environments

A. E. V. Olmos (Corresponding Author) and E. P. G. Pinto are with Universidad Militar Nueva Granada, Bogotá, Colombia. Emails: edward.guillen@unimilitar.edu.co and est.andres.villama1@unimilitar.edu.co

with strict operational constraints, such as edge nodes or IoT gateways.

This paper is organized as follows: Section 2 reviews the state of the art regarding IDSs and feature-engineering techniques. Section 3 describes the UNSW-NB15 dataset and its statistical distribution. Section 4 details the methodology, including the pipeline for packet labeling, traffic grouping, feature extraction and the experimental evaluation framework. Section 5 discusses the results obtained from the comparative analysis between the official and the reconstructed dataset. Finally, Section 6 summarizes the main conclusions and outlines directions for future research.

2. RELATED WORK

Network exposure to attack continues to increase and evolve over time, resulting in the constant emergence of previously unseen variants and attack methods. Consequently, a significant proportion of the current literature evaluates its methodologies using the UNSW-NB15 dataset, which integrates real traffic with updated malicious traffic, provides 49 attributes and covers nine contemporary families of attacks, making it a more representative tool compared to historical datasets, such as KDD-99 [3]. On this basis, the works are grouped into three lines: (i) classical learning with feature selection, (ii) deep learning, often preceded by dimension reduction and (iii) deployment-oriented ensembles/hybrids. Table 1 summarizes the key methodologies and reported performance metrics across these research lines.

Classical learning with feature selection. First, across the following studies, researchers combine supervised algorithms with feature selection to reduce dimensionality, mitigate overfitting and decrease computational cost. Methodologically, these studies typically begin with exploratory and correlation analyses that assess attribute relevance and eliminate redundancies before modeling. More et al. [4] strengthen classical models on UNSW-NB15 by cleaning the dataset, running exploratory and correlation analyses, estimating attribute relevance (e.g., with XGBoost) and removing redundant variables. The authors then compare Logistic Regression, SVM, Decision Tree and Random Forest to select the best detector. The study finds that, after feature selection, Random Forest consistently outperforms the alternatives, underscoring the impact of variable filtering on performance.

Ahmad et al. [5] extend the attribute-prioritization approach to an IoT setting aligned with protocol specifics. The authors group features by domain (Flow/MQTT and TCP) and exclude features that induce overfitting, then evaluate RF, SVM and Neural Networks per cluster and in combination. The authors report that the cluster strategy maintains high effectiveness with Random Forest while reducing training time relative to general supervised pipelines, demonstrating the efficacy of protocol-sensitive filtering. Hussain et al. [6] analyze detection under a Zero Trust framework with continuous network operation. The authors prepare the data *via* imputation, encoding and normalization, apply Recursive Feature Elimination (RFE) to identify predictive variables and compare Logistic Regression, Random Forest and XGBoost. XGBoost achieves an AUC of 1.00, indicating strong capacity to capture non-linear relationships in network traffic.

Deep learning, with dimension reduction. Another significant line of research employs deep networks together with dimension reduction to accelerate inference and improve generalization, particularly in resource-constrained scenarios (e.g., IoT) where classical models face limitations in complex, high-volume settings. Jouhari et al. [7] design an efficient IoT IDS using a lightweight CNN-BiLSTM model. A Chi-square selection process reduces the input to the 20 most relevant features before training on UNSW-NB15. The design aims for high accuracy with low complexity; the reported performance is 97.90% (binary) and 97.09% (multi-class), with lower prediction latency attributable to feature reduction.

Sharma and Kumar [8] propose a CapsNet+BiLSTM hybrid that exploits spatial hierarchies and temporal dependencies. Capsules with dynamic routing replace classical convolutions to extract local features and preserve hierarchical information, thereby reducing dimensionality and enhancing feature representation. The architecture then adds BiLSTM for bidirectional sequence modeling to capture temporal dependencies. The authors evaluate the model on CIC-IDS2017, KDD CUP 99 and UNSW-NB15, reporting improvements over individual architectures and 97% accuracy on UNSW-NB15. Vibhute et al. [9] combine feature selection with deep learning by selecting 15 of the 49 UNSW-NB15 attributes *via* Random Forest and training a CNN on the reduced sub-set. The pipeline aims to simplify

the input space and improve generalization, achieving 99.00% test accuracy, 98.86% recall and 99.00% F1. By contrast, Farhan et al. [10] build a sequential DNN preceded by Extra Trees feature selection, reducing 43 features to 8. The reduction improves computational efficiency and inference speed while maintaining effective attack discrimination; on UNSW-NB15 (binary), the authors report 97.93% accuracy and 97% recall/F1.

Deployment-oriented ensembles/hybrids. Finally, transitioning towards production environments, several studies emphasize ensemble and hybrid systems that balance accuracy, latency and robustness, often with feature selection and distributed execution. Chkirbene et al. [11] exemplify this direction with a hybrid that selects significant features using Random Forest and classifies attacks with CART. CART scales well to large datasets and adapts tree structure to input variables. Compared with alternative trees and baselines, the hybrid improves accuracy while reducing complexity and training/prediction time. Kabir et al. [12] study stacking for better generalization using two configurations: (i) XGBoost and KNN as base models with Random Forest as meta-classifier and (ii) XGBoost, Neural Networks and KNN with the same meta-classifier. The pre-selection phase uses Extra Trees and Mutual Information Gain. On UNSW-NB15, the combination of Mutual Information Gain with the first stacking configuration (XGBoost+KNN) reaches 96.24% accuracy, surpassing individual models and highlighting the value of heterogeneous ensembles for complex attack patterns. With a deployment and scalability focus, Belouch et al. [13] implement an Apache Spark pipeline and evaluate SVM, Naïve Bayes, Decision Tree and Random Forest on UNSW-NB15. Random Forest outperforms the other classifiers, achieving 97.49% accuracy with 0.08 s inference time.

Furthermore, Mutambik [14] proposed IoT-FIDS (Flow-based Intrusion Detection System for IoT), with a focus on reducing computational complexity in resource-constrained networks. Unlike methods based on computationally expensive traditional ML algorithms, IoT-FIDS identifies anomalous behaviors by analyzing flow-based representations that capture communication patterns. This use of flow-level features is key to reducing dimensionality and data volume, thereby justifying the adoption of flow-based representations to achieve a balance between efficiency and accuracy in IoT gateways, as proposed in our study.

Table 1. Metrics reported in Related Work on UNSW-NB15.

Reference	Model / Approach	Reported Metrics
[7]	CNN-BiLSTM (Chi-square: 20 features)	ACC = 97.90%(binary); ACC = 97.09%(multi)
[8]	CapsNet + BiLSTM	ACC = 97.00% (UNSW-NB15)
[9]	Random Forest feature selection (15/49) + CNN	ACC = 99.00%; Recall = 98.86%; F1-score = 99.00%
[10]	Extra Trees (8 feats) + DNN	ACC = 97.93%(binary); Recall = 97.00%; F1-score = 97.00%
[12]	Stacking: XGBoost + KNN → Random Forest (with MI gain)	ACC = 96.24%
[13]	Random Forest (Apache Spark)	ACC = 97.49%; Latency = 0.08 s

Note: The work by Mutambik [14] is excluded from this comparison, since its performance metrics were evaluated on the BoT-IoT dataset, utilizing UNSW-NB15 exclusively for training.

3. DATASET UNSW-NB15

UNSW-NB15 is a widely recognized benchmark representing generic enterprise network traffic. It was generated at the UNSW Canberra laboratory with normal and malicious simulated traffic from contemporary attacks. The captures were carried out over two sessions: on 22-01-2015 with a duration of 16 hours and on 17-02-2015 with a duration of 15 hours, totaling 99.1 GB of PCAP files. From these PCAPs, the authors extracted features using Argus and Bro-IDS to produce the official CSVs with 49 features and 9 attack categories, the breakdown of which by class is presented in Table 2.

In this study, we process all 93,715,272 raw packets from the 15 -hour capture PCAP files to build our custom CSV. This choice is based on the better balance between the normal and attack classes compared to the 16 -hour capture session. According to the figures in Table 3, in the 16 -hour capture, the attacks account for only 2.04% (22,215 attacks out of 1,087,202 total records), whereas the 15-hour capture reaches 20.59% (299,068 attacks out of 1,452,842 total records). This indicates that, although using the

15 -hour capture significantly increases the representation of attack categories, a serious class imbalance persists throughout the dataset. For example, in Table 2, while the Normal category comprises 2,218,761 records (87.35%), minority classes such as Worms represent only 174 records (0.007%), posing a significant challenge for detection stability and model training.

Table 2. Distribution of the UNSW-NB15 dataset.

Category	**No. Records**
Normal	2,218,761
Fuzzers	24,246
Analysis	2,677
Backdoors	2,329
DoS	16,353
Exploits	44,525
Generic	215,481
Reconnaissance	13,987
Shellcode	1,511
Worms	174

Table 3. Statistics per day of capture at UNSW-NB15.

Feature	22-01-15 (16h)	17-02-15 (15h)
Src_bytes	4,860,168,866	5,940,523,728
Dst_bytes	44,743,560,943	44,303,195,509
Src_Pkts	41,168,425	41,129,810
Dst_pkts	53,402,915	52,585,462
Normal records	1,064,987	1,153,774
Attack records	22,215	299,068

The information presented in Tables 2 and 3 was obtained directly from the work [3] by Moustafa et al.

4. METHODOLOGY

We followed a four-stage pipeline: (i) label each packet with its category; (ii) group the packets into unidirectional flows; (iii) compute features per flow and consolidate a single CSV; and (iv) apply pre-processing, training and evaluation of machine-learning models, comparing the own CSV (from PCAP) against the official CSVs, as illustrated in Figure 1.

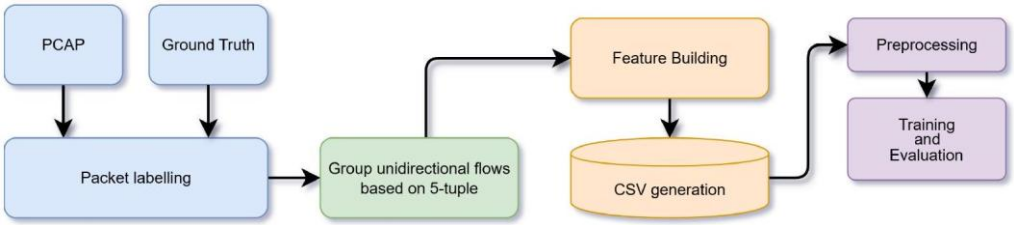


Figure 1. Methodology pipeline.

4.1 Packet Labeling

In addition to the raw PCAP files available in the UNSW-NB15 repository, the set includes a ground truth file that contains key information: the start and end time of each attack in Unix time, its category, the protocol and the source and destination IP addresses and ports.

Based on this file, each packet in the PCAPs was labeled as normal or as one of the nine attack families. To do this, from each packet its timestamp (Unix time) and IP addresses were extracted and then contrasted with the ⟨Source IP, Destination IP, Start time, Last time⟩ records of the ground truth. If

the packet's timestamp fell within the interval [Start time, Last time] associated with that pair of IPs, the corresponding Attack category from the ground truth was assigned; otherwise, it was labeled as normal, since the ground truth only records malicious traffic events. The procedure is summarized in Algorithm 1.

Algorithm 1: Packet Labeling using Ground Truth

```
Data: PCAP packets; GroundTruth with (SrcIP, DstIP, StartTime, LastTime,
      AttackCategory)
Result: Label per packet (normal or attack category)
foreach packet p in PCAP do
    PacketTimestamp ← timestamp(p);
    (PacketSrcIP, PacketDstIP) ← (IP.src(p), IP.dst(p));
    Label ← normal;
    foreach record r in GroundTruth with r.SrcIP = PacketSrcIP and r.DstIP =
     PacketDstIP do
        if r.StartTime ≤ PacketTimestamp ≤ r.LastTime then
            Label ← r.AttackCategory; break;
    save Label for p;
```

4.2 Flow Grouping

With the packets labeled, we proceed to the creation of unidirectional flows. To this end, a grouping strategy based on the 5-tuple method is proposed, together with three flow-closing strategies: category change, a time threshold without meeting the condition and, finally, exceeding a maximum lifetime for an active flow.

To identify the packets belonging to the same unidirectional flow, the 5 -tuple method is used, defined as [source IP, destination IP, source port, destination port, protocol]. The process consists of iterating through the packets one by one in temporal order and grouping those that maintain the same 5 -tuple values, which indicates that they belong to the same unidirectional flow.

The flow must be closed when any of the following three conditions is met: (i) a new packet does not belong to the same category as the other packets in the flow, even if it matches the 5 -tuple; (ii) the flow does not receive new packets that satisfy the 5 -tuple for 15 seconds; or (iii) the flow reaches a maximum active lifetime of 60 seconds from its first packet (even if it remains active by adding more packets).

The selection of these timeout parameters is aligned with established network industry standards and operational monitoring guidelines. The 15 -second idle timeout is a widely adopted configuration in Cisco NetFlow environments and conforms to the IPFIX protocol specifications defined in RFC 5101. Regarding the active timeout, while standard enterprise implementations may employ longer intervals of 30 minutes to minimize processing overhead, RFC 5101 specifies that long-lived flows should be periodically exported to ensure continuous visibility. In the context of intrusion detection, reducing this threshold to 60 seconds is essential for enabling segmented traffic analysis and facilitating near real-time threat response. This design choice is consistent with prominent flow-based extraction frameworks in the literature, such as CICFlowMeter, which utilizes a 60 -second limit to prevent delayed detection of persistent anomalous behaviors. The procedure for creating and closing flows is summarized in Algorithm 2.

It is important to highlight that both the packet-labeling mechanism (Algorithm 1) and the subsequent flow-grouping stage (Algorithm 2) operate under the assumption of static IP allocations and the absence of overlapping, NATed (Network Address Translation) flows. In real-world environments where multiple internal devices share a single public IP address, concurrent sessions may overlap temporally or collide within the same 5-tuple structure, representing an inherent limitation of this dataset construction framework.

4.3 Feature Calculation

To justify the feature design selection, a two-stage workflow was implemented, combining statistical

filtering and explaining ability validation. Initially, a broad set of candidate features was extracted from the raw unidirectional flows. To reduce dimensionality and eliminate uninformative variables, a selection phase based on Mutual Information (MI) was performed. As shown in Figure 2a, features with an importance score above 1% (MI>0.01) were retained, discarding attributes with low predictive power.

Algorithm 2: Unidirectional Flow Construction (5-tuple and closure rules)

Data: LabeledPackets (time-ordered), each with {Timestamp, SrcIP, DstIP, SrcPort, DstPort, Protocol, Category}
Result: Set of unidirectional flows grouped by 5-tuple and Category
IDLE_TIMEOUT ← 15 s; *MAX_LIFETIME* ← 60 s;
ActiveFlows ← empty map from 5-tuple to flow state;
foreach *packet p in LabeledPackets* **do**
  key ← (*p.SrcIP*, *p.DstIP*, *p.SrcPort*, *p.DstPort*, *p.Protocol*);
  if *key* ∉ *ActiveFlows* **then**
    open new flow with *Category* ← *p.Category*, *StartTime* = *LastTime* = *p.Timestamp*;
    ActiveFlows[*key*] ← flow;
    continue;
  flow ← *ActiveFlows*[*key*];
  if *p.Category* ≠ *flow.Category* **or** (*p.Timestamp* − *flow.LastTime*) ≥ *IDLE_TIMEOUT* **or** (*p.Timestamp* − *flow.StartTime*) ≥ *MAX_LIFETIME* **then**
    emit flow;;
    remove *ActiveFlows*[*key*];
    open new flow with *Category* ← *p.Category*, *StartTime* = *LastTime* = *p.Timestamp*;
    ActiveFlows[*key*] ← flow;
  else
    flow.*LastTime* ← *p.Timestamp*;
foreach *remaining flow in ActiveFlows* **do**
  emit flow

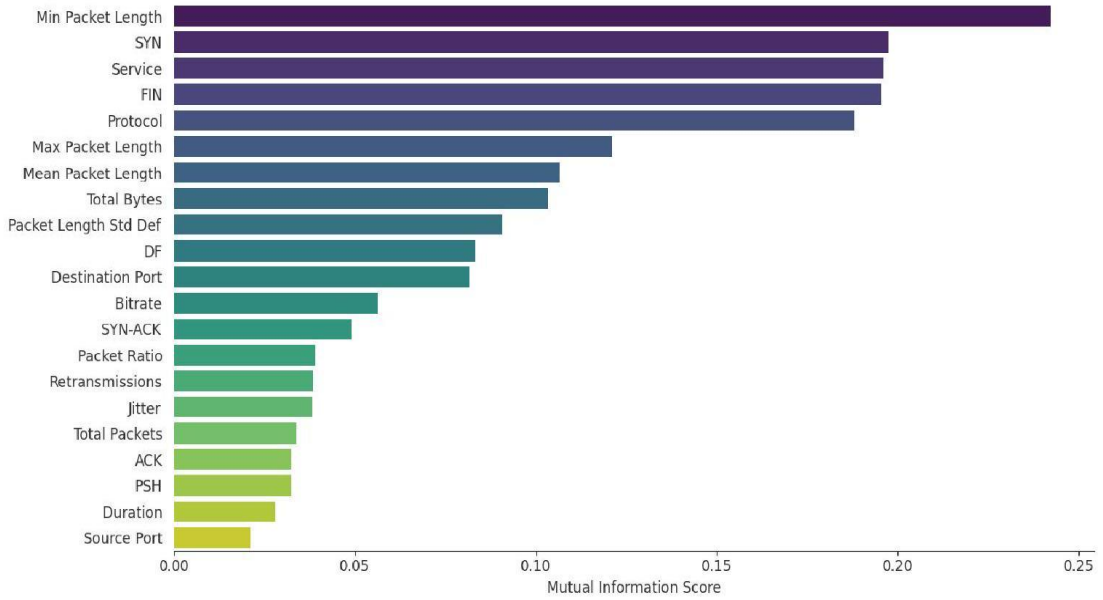
Subsequently, to validate this selection from the model's perspective, a SHAP (SHapley Additive exPlanations) analysis was conducted using a XGBoost model for both binary and multi-class classification tasks (Figure 2b and Figure 2c). This analysis confirmed the individual contribution of each variable. For instance, it was found that attributes statistically significant in Mutual Information, such as the SYN flag, were also important for the binary XGBoost model, but showed less relevance in multi-class classification. This demonstrates how attribute importance varies depending on the predictive goal: while some features serve as general anomaly indicators (attack *vs.* normal traffic), others are required to distinguish between specific attack signatures. The result is a refined set of 21 features (detailed in Table 4) that reduce training complexity and dataset size without compromising detection accuracy.

The 21 features were computed over the packets of each flow and stored in a single CSV file. The generated CSV file contains a total of 2,064,494 records corresponding to the features of each flow, resulting in a final size of 220 MB, which is 2.5 times smaller compared to the total of 560 MB of the official CSV files.

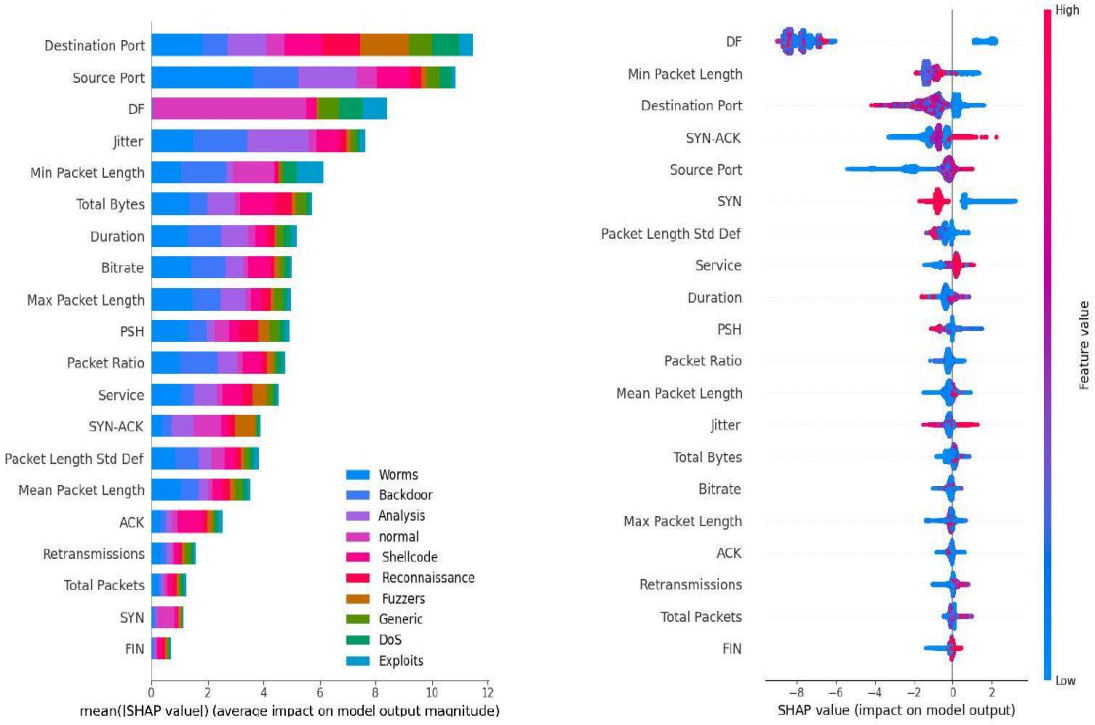
4.4 Pre-processing and Evaluation

To ensure a fair comparison, the full CSV version of the UNSW-NB15 dataset was employed. However, a feature filtering process was implemented to align this set with the attributes present in the official train/test partitions. This decision was made to maintain methodological consistency and, crucially, to

prevent data leakage. Beyond this feature alignment, an identical pre-processing pipeline was applied to both the official UNSW-NB15 CSV and our custom representation derived from the raw PCAP files.



(a) Mutual Information scores.



(b) SHAP importance (Multi-class). (c) SHAP importance (Binary).

Figure 2. Feature analysis: (a) Mutual Information scores for the selected features set; (b) and (c) SHAP global importance for model explainability.

Pre-processing consists of the following:

- Removal of duplicate records and/or records with missing values (NaN).
- Numerical coding of labels.
- Random mix of records.
- Stratified 5-fold Cross-validation protocol.

Table 4. Unidirectional 21 flow features: definition and unit.

Feature	Definition	Unit
Total Packets	Total number of packets in the flow	packets
Source Port	Source transport-layer port number	integer
Destination Port	Destination transport-layer port number	integer
Protocol	IP protocol identifier	integer
Duration	Elapsed time between first and last packet	s
Min. Packet Length	Smallest packet size observed in the flow	bytes
Max. Packet Length	Largest packet size observed in the flow	bytes
Mean Packet Length	Average packet size in the flow	bytes
Packet Length Std. Dev.	Standard deviation of packet sizes	bytes
Total Bytes	Sum of all packet lengths in the flow	Bytes
DF	'Don't Fragment' bit in the IP header	binary
FIN	TCP 'Finish' flag	binary
SYN	TCP 'Synchronize' flag	binary
PSH	TCP 'Push' flag	binary
ACK	TCP 'Acknowledgment' flag	binary
Retransmissions	Count of retransmitted segments within the flow	integer
Service	Service or application name associated with the observed port	string
Bitrate	Average number of bits transmitted per second over the flow	bps
Jitter	Average variability of inter-packet arrival times	s
SYN-ACK	Delay from the initial SYN to the corresponding SYN-ACK	s
Packet Ratio	Number of packets transmitted in a second	integer

Since our methodology calculates feature flows derived from the raw packets using grouping, calculation and structural rules that differ from the official dataset, it is impossible to directly employ the official test split for a fair evaluation. By using stratified 5-fold cross-validation, we ensure an identical and unbiased evaluation environment for both datasets.

The label was handled under two tasks: binary (normal *vs.* attack) and multi-class (normal +9 attack families) across both data sources. We selected four models per task (Decision Tree, Random Forest, MLP and XGBoost), for a total of 16 training experiments (8 per CSV). Each model and its configuration are summarized in Table 5. For the MLP, only the output layer is adjusted according to the task (sigmoid for binary; softmax for multi-class).

Table 5. Training configuration for each model.

Model	Configuration for binary and multi-class
Decision Tree	Minimum samples split: 2; minimum samples per leaf: 1; criterion: Gini; class weight: balanced.
Random Forest	Trees: 100; minimum samples split: 2; minimum samples per leaf: 1; criterion: Gini; class weight: balanced.
MLP	Layers: Dense(512, ReLU) → Dropout(0.3) → Dense(256, ReLU) → Dense(n classes, Sigmoid/Softmax). Optimizer: Adam (Initial LR = 0.001, exponential decay to 0.0001 from epoch 50); epochs: 150; batch size: 8192 ; class weight: balanced.
XGBoost	Trees: 1000; max depth: 8; LR: 0.1 (exponential decay to 0.01 from round 100); subsample/colsample: 0.6; regularization: gamma = 0.1, alpha = 0.5, lambda = 1.5; sample weight: balanced.

The models were evaluated using a Stratified 5-fold Cross-validation scheme to ensure robustness and

account for class imbalance. For each fold, a comprehensive set of performance and computational efficiency metrics was recorded; the results across all folds are presented and discussed in Section 5.

4.5 Statistical Significance

To validate the performance evaluations and determine whether the observed differences between models and datasets are statistically significant, two non-parametric tests were employed. This statistical validation is essential in network-intrusion detection to ensure that performance claims are robust, especially when dealing with the inherent class imbalance of traffic datasets.

4.5.1 McNemar's Test

McNemar's test was employed to perform pairwise comparisons between the classification models (Decision Tree, Random Forest, MLP and XGBoost) on the PCAP-derived dataset. This test is particularly suited for IDS evaluation as it focuses on discordant prediction cases where one model classifies an attack correctly while the other fails. By analyzing these prediction shifts, the test evaluates the null hypothesis that both classifiers have an equal proportion of errors, providing a more rigorous basis for asserting model superiority in imbalanced scenarios.

4.5.2 Wilcoxon Signed-rank Test

While McNemar's test evaluates model performance on a single dataset, the Wilcoxon Signed-rank test was implemented to compare the proposed PCAP-derived representation against the original UNSW-NB15 dataset. Using XGBoost as the benchmark architecture, this test evaluates whether the performance shift between datasets is consistent across 10 folds of Stratified K-fold cross-validation. This 10 -fold configuration was specifically chosen for the statistical comparison to provide a more robust sample size, distinct from the 5 -fold setup used during the initial training phase. Unlike parametric tests, Wilcoxon does not assume a normal distribution of the results, making it ideal for comparing metrics, such as latency, training time and F1-score across different data representations.

5. RESULTS AND DISCUSSION

Table 6 summarizes the performance of each model using a Stratified 5-fold Cross-validation evaluation scheme. The mean values for each metric are presented alongside their respective standard deviations for both datasets in binary and multi-class classification tasks. As expected, the task of discriminating between normal traffic and attacks presents an intrinsically lower complexity than identifying specific attack categories; consequently, performance metrics in binary classification are significantly higher across all evaluated models. It is evident that the proposed PCAP-derived approach, by reducing the dimensionality of the input data, optimizes inference time and enables the processing of a higher volume of samples per second-as measured on an NVIDIA A100-SXM4-40GB GPU-in exchange for a minor reduction in detection capability. For instance, the XGBoost model, which demonstrated statistical superiority over the other models, experienced a 6% decrease in its F1-score for multi-class classification, offset by a 48% increase in samples per second. In the binary scenario, a 6% reduction in the F1-score was also observed, against a substantial 68.8% increase in processing performance.

Notably, tree-based models, such as Decision Tree, Random Forest and XGBoost, exhibit an increase in the number of parameters despite being trained on a reduced feature set. This phenomenon suggests that these models require greater depth and node density to compensate for the reduced dimensional information and maintain classification accuracy. However, this structural complexity does not correlate directly with inference speed; it is observed that models with a higher parameter count can achieve superior processing speeds. In contrast to the MLP model, the results confirm that the relationship between parameter load and computational efficiency is not universal, but depends strictly on the nature of the algorithm.

The results of the McNemar test for all pairwise model comparisons are presented in Table 7. Across both binary and multi-class tasks, all comparisons yielded p -values under 0.001, confirming that the performance differences between the models are statistically significant.

The highest test statistics are consistently associated with the MLP model, particularly when compared against Random Forest and XGBoost. This high divergence indicates that the MLP produces a

substantially different error distribution, likely due to its neural network architecture failing on different types of traffic compared to the tree-based ensembles. In contrast, the comparison between Decision Tree and XGBoost shows the lowest statistical divergence. This suggests a closer logical alignment between these two models; however, the test still confirms that the two classifiers produce distinct results, validating that the performance gap observed between the baseline Decision Tree and the optimized XGBoost is statistically significant.

Table 6. Evaluation results for each model.

Model	Metric	Multi-class		Binary	
		PCAP-derived CSV	UNSW-NB15 CSV	PCAP-derived CSV	UNSW-NB15 CSV
Decision Tree	ACC	$0.977 \pm 2e-4$	$0.98 \pm 3e-4$	$0.984 \pm 2e-4$	$0.99 \pm 1e-4$
	AUC	0.727 ± 0.008	0.776 ± 0.003	0.86 ± 0.002	$0.943 \pm 7e-4$
	Precision*	0.477 ± 0.017	0.562 ± 0.007	0.872 ± 0.002	$0.95 \pm 8e-4$
	Recall*	0.481 ± 0.016	0.542 ± 0.005	0.86 ± 0.002	$0.943 \pm 7e-4$
	F1-score*	0.478 ± 0.014	0.548 ± 0.02	0.866 ± 0.002	$0.946 \pm 7e-4$
	Param.	$96.3\ K \pm 271$	$55.5\ K \pm 641$	$73.5\ K \pm 1.41\ K$	$35.8\ K \pm 850$
	Sample/s	$2.51M \pm 69\ K$	$1.99M \pm 58.4\ K$	$2.9M \pm 157\ K$	$2.15M \pm 94.3\ K$
Random Forest	ACC	$0.982 \pm 1e-4$	$0.983 \pm 1e-4$	$0.986 \pm 1e-4$	$0.993 \pm 1e-4$
	AUC	0.946 ± 0.007	0.936 ± 0.002	$0.994 \pm 3e-4$	$0.999 \pm 1e-4$
	Precision*	0.608 ± 0.017	0.619 ± 0.01	$0.893 \pm 7e-4$	$0.966 \pm 4e-4$
	Recall*	0.502 ± 0.006	0.557 ± 0.007	0.883 ± 0.002	$0.955 \pm 9e-4$
	F1-score*	0.537 ± 0.008	0.579 ± 0.007	0.888 ± 0.001	$0.961 \pm 6e-4$
	Param.	$5.24M \pm 15.1\ K$	$3.78M \pm 11\ K$	$3.69M \pm 8.91\ K$	$2.15M \pm 8.79\ K$
	Sample/s	$280\ K \pm 9.8\ K$	$240\ K \pm 5.46\ K$	$537\ K \pm 24\ K$	$569\ K \pm 21.5\ K$
MLP	ACC	0.956 ± 0.002	$0.972 \pm 5e-4$	$0.977 \pm 6e-4$	$0.986 \pm 4e-4$
	AUC	0.991 ± 0.001	$0.997 \pm 1e-4$	$0.996 \pm 2e-4$	$0.999 \pm 1e-4$
	Precision*	0.316 ± 0.005	0.487 ± 0.006	0.789 ± 0.003	0.888 ± 0.002
	Recall*	0.695 ± 0.008	0.706 ± 0.009	$0.986 \pm 4e-4$	$0.992 \pm 2e-4$
	F1-score*	0.355 ± 0.008	0.516 ± 0.008	0.86 ± 0.003	0.933 ± 0.002
	Param.	145.2 K	155.4 K	142.8 K	153.1 K
	Sample/s	$19.8\ K \pm 2.07\ K$	$19.8\ K \pm 1.68\ K$	$21.4\ K \pm 270$	$21\ K \pm 152$
XGBoost	ACC	$0.975 \pm 2e-4$	$0.977 \pm 3e-4$	$0.982 \pm 1e-4$	$0.989 \pm 2e-4$
	AUC	$0.996 \pm 47e-4$	$0.997 \pm 1e-4$	$0.997 \pm < 1e-4$	$0.999 \pm < 1e-4$
	Precision*	0.481 ± 0.014	0.584 ± 0.008	$0.823 \pm 8e-4$	0.908 ± 0.001
	Recall*	0.674 ± 0.009	0.669 ± 0.007	$0.984 \pm 3e-4$	$0.99 \pm 4e-4$
	F1-score*	0.547 ± 0.009	0.607 ± 0.007	$0.885 \pm 7e-4$	$0.945 \pm 8e-4$
	Param.	$1.93M \pm 7.48\ K$	$1.68M \pm 17.9\ K$	$263\ K \pm 2.41\ K$	$243\ K \pm 2.74\ K$
	Sample/s	$364\ K \pm 9.7\ K$	$246\ K \pm 4\ K$	$1.43M \pm 79.4\ K$	$847K \pm 42.1\ K$

*The Precision, Recall and F1-score metrics are calculated as macro averages, executed on A100 GPU.

Table 7. McNemar test results for model-performance comparison.

Comparison	Multi-class		Binary	
	Statistic	p-value	Statistic	p-value
Decision Tree *vs.* Random Forest	3,158.27	< 0.001	1,439.14	< 0.001
Decision Tree *vs.* MLP	26,816.13	< 0.001	4,090.86	< 0.001
Decision Tree *vs.* XGBoost	863.23	< 0.001	218.52	< 0.001
Random Forest *vs.* MLP	36,677.67	< 0.001	7,897.61	< 0.001
Random Forest *vs.* XGBoost	5,241.43	< 0.001	1,895.01	< 0.001
MLP *vs.* XGBoost	34,287.12	< 0.001	9,397.44	< 0.001

The statistical comparison between the proposed PCAP-derived dataset and the original UNSW-NB15 representation, conducted using XGBoost across 10 folds of Stratified K-fold cross-validation, is summarized in Table 8. The Wilcoxon Signed-rank test confirms that the performance variations between both datasets are statistically significant ($p = 0.0019$) across all evaluated metrics, validating that these differences are consistent and not a result of stochastic variance.

Table 8. Wilcoxon Signed-rank test results comparing PCAP-derived and UNSW-NB15 datasets' performance across metrics.

Metric		Multi-class			Binary		
		PCAP-derived	UNSW-NB15	p-value	PCAP-derived	UNSW-NB15	p-value
Accuracy		0.9746	0.9766	0.0019	0.9822	0.9887	0.0019
AUC		0.9962	0.9973	0.0019	0.9966	0.9993	0.0019
F1 Macro		0.5474	0.6073	0.0019	0.8852	0.9442	0.0019
T4	Train Time (s)	180.0457	236.9737	0.0019	20.8764	28.3762	0.0019
	Latency (ms)	0.0142	0.0115	0.0019	0.0014	0.0024	0.0019
A100	Train Time (s)	62.7990	77.8722	0.0019	11.1072	13.6984	0.0019
	Latency (ms)	0.0036	0.0052	0.0019	0.0008	0.0013	0.0019

* Hardware platforms are denoted as T4 (NVIDIA Tesla T4 GPU) and A100 (NVIDIA A100-SXM4-40GB GPU).

As previously noted, the original UNSW-NB15 dataset maintains a slight advantage in detection metrics. Specifically, Accuracy and AUC differences remain within approximately 0.2% and 0.1% for multi-class tasks and 0.6% and 0.3% for binary classification, respectively. While a more noticeable trade-off of approximately 6% is observed in the F1-score, the PCAP-derived representation offers substantial gains in computational efficiency. On the T4 GPU architecture, training time was reduced by 24.0% for multi-class and 26.4% for binary classification, a trend that persists on the A100 architecture with reductions of 19.3% and 18.9%, respectively.

A notable exception occurs in the multi-class inference latency on the T4 platform, where the PCAP-derived dataset exhibited higher latency (0.0142 ms) compared to the original set (0.0115 ms). This behavior is attributed to the increased structural complexity of the model: to compensate for the reduced feature space, the XGBoost algorithm generated deeper trees to capture underlying traffic patterns. Specifically, the PCAP-derived model required 2,008,476 total nodes, a 7.3% increase over the 1,871,454 nodes produced by the official dataset. This higher node density leads to increased memory pressure and cache misses on the T4's memory hierarchy, which has more limited resources. However, this effect is mitigated on the A100 platform; its significantly larger L2 cache and superior memory bandwidth allow for more efficient traversal of complex tree structures, ultimately achieving latency reductions of up to 38% in binary tasks.

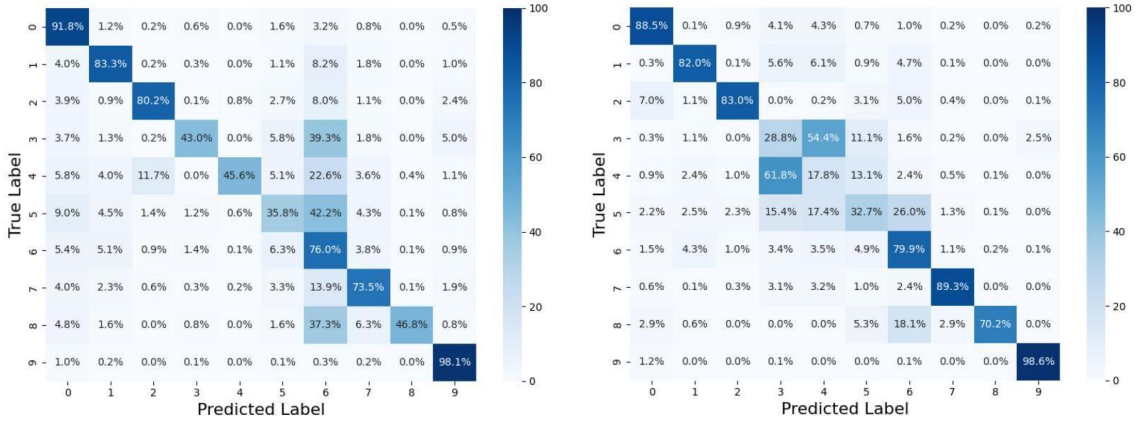
Compared with the studies in Related Work in Table 1, our approach is competitive at a lower computational cost. This is explained by the smaller number of classes and the reduced size of PCAP-derived CSV relative to the official datasets, without significantly sacrificing performance.

Additionally, the confusion matrix for the XGBoost model evaluated on the PCAP-derived dataset (Figure 3a) exhibits a clearer diagonal trend across several categories, indicating a redistribution of classification performance compared to the original representation. However, a detailed analysis of class 6 (Exploits) reveals a performance trade-off: while the model demonstrates a moderate recall of 76% in identifying this specific class, its precision is compromised due to misclassifications originating from other categories, primarily class 3 (Analysis), 5 (DoS) and 8 (Worms).

In contrast, the official UNSW-NB15 dataset (Figure 3b) displays significant overlap across several classes and substantially higher false positive rates. For instance, class 4 (Backdoor) is mislabeled as class 3 at a rate of 61.8% and conversely, class 3 is also mislabeled as class 4 in 54.4% of cases. Overall, classes 3, 4 and 5 (Analysis, Backdoor and DoS, respectively) exhibit the most critical false-positive rates, as the model largely fails to discriminate between them.

In Figure 3a, class 6, representing malicious Exploit traffic, acts as the primary source of ambiguity for the multi-class classifier. This category induces significant noise, elevating the overall false-positive

rate and diminishing the model's discriminatory power across the dataset. Given this pattern, it may be advisable to evaluate the exclusion of this class from the general multi-class scheme or its isolation into an independent detector to strictly control false-positive rates. Although such a modular architecture increases computational overhead and operational complexity compared to a single multi-class classifier, this trade-off is justifiable when specific class overlaps significantly degrade the system's reliability.



(a) Confusion matrix (PCAP-derived). (b) Confusion matrix (UNSW-NB15).

Figure 3. Confusion matrix for the XGBoost multi-class classifier: (a) Evaluated on PCAP-derived CSV; (b) Evaluated on UNSW-NB15 CSV. Class indices: 0: Fuzzers, 1: Reconnaissance, 2: Shellcode, 3: Analysis, 4: Backdoor, 5: Dos, 6: Exploits, 7: Generic, 8: Worms, 9: Normal.

The per-class F1-score, illustrated in Figure 4, provides a more comprehensive assessment of the model's ability to balance precision and recall across the unbalanced categories of the UNSW-NB15 dataset. The results reveal a complex trade-off: the PCAP-derived representation significantly improves performance in the most critical minority classes where the official dataset shows its worst results. Notably, the F1-score for Backdoor increased from a negligible 0.09 to 0.45 and Analysis rose from 0.12 to 0.20 , suggesting that the reduced feature set captures more effective patterns for these specific threats.

Conversely, the reduction in features leads to a performance decline in other categories. The most substantial drop is observed in the Generic class, where the F1-score decreased from 0.93 to 0.52, alongside moderate reductions in Reconnaissance (0.85 to 0.68), Exploits (0.82 to 0.68) and Worms (0.66 to 0.45). This indicates that while the custom PCAP-derived with 21 feature set mitigates the total failure of the model to detect certain stealthy attacks, it also sacrifices discriminative power in classes that were well-supported by the original UNSW-NB15 features.

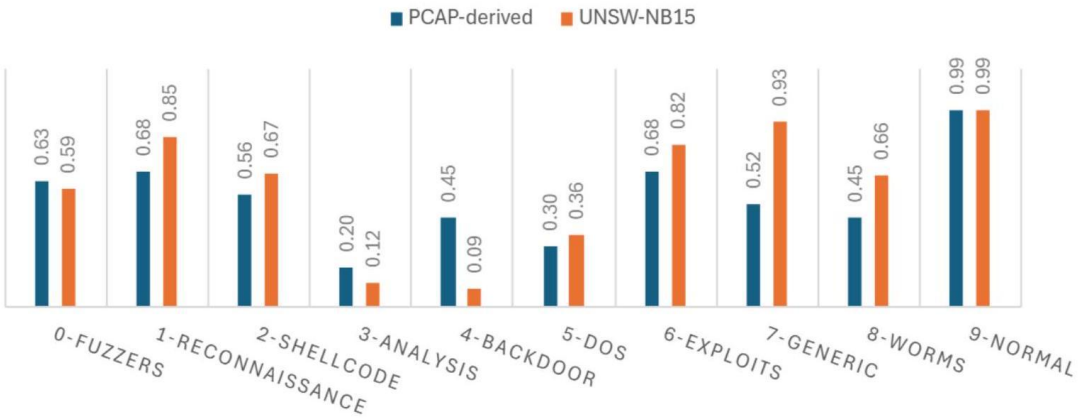


Figure 4. Bar chart showing per-class $f1$-score values for XGBoost model.

The performance of the proposed method compared to other recent works using the UNSW-NB15 dataset is presented in Table 9. This comparison focuses on accuracy metrics for both binary and multi-class tasks, where the proposed approach reports values of 98.60% and 98.20%, respectively. It is

important to clarify that these specific metrics correspond to the results achieved by the Random Forest model, which demonstrated the highest overall accuracy among all evaluated architectures in this study; however, this model is not necessarily the most suitable choice for scenarios requiring a balanced detection of minority classes. As previously discussed, although the Random Forest leads in raw accuracy, models like XGBoost provide a more robust and equitable performance across the entire attack spectrum.

Table 9. Performance comparison of the proposed approach with recent literature on the UNSW-NB15 dataset.

Reference	Accuracy (ACC)	
	Binary	Multi-class
[7]	97.90%	97.09%
[8]	97.00%	-
[9]	99.00%	-
[10]	97.93%	-
[12]	96.24%	-
[13]	97.49%	-
Proposed Work	98.60%	98.20%

The metrics obtained by our approach are highly competitive with recent literature. However, it is important to note that direct numerical comparisons with these state-of-the-art works serve as a contextual reference rather than an absolute benchmark, due to differences in evaluation protocols. While the referenced studies evaluate their models using the static, down-sampled official UNSW-NB15 test partition, our methodology necessitated a stratified 5-fold cross-validation on the full PCAP-derived capture to ensure an unbiased evaluation of our custom-flow construction. Although the main focus of this research was not to exhaustively maximize classification performance, but rather to evaluate the effectiveness of a reduced feature set within a lightweight and computationally efficient framework, the outcomes are highly satisfactory. The custom representation proves that it is possible to achieve state-of-the-art intrusion-detection accuracy while maintaining a reduced profile suitable for resource-constrained networks.

6. CONCLUSIONS

This study quantitatively evaluated a lightweight, network flow-based data representation constructed from the raw PCAP files of the UNSW-NB15 dataset, reducing the feature space from 49 to 21 attributes and achieving a file size 2.5 times smaller than the original. Experimental results demonstrated that this dimensional reduction offers an optimal trade-off for lightweight IDS deployments; the statistically superior XGBoost model experienced a minor 6% macro F_1-score reduction across both tasks, which was heavily compensated for by reduced training times and substantial inference speed increases of 68.8% in binary and 48% in multi-class classification.

Furthermore, this investigation revealed that the detection rate for specific attack classes varies significantly depending on how the dataset features are computed. Experimental results proved that certain features are highly effective for detecting specific attack categories, but less favorable for others. Quantitatively, this performance disparity is particularly evident in minority classes, such as Analysis and Backdoor, where the proposed methodology successfully increased the recall rates from 29% to 43% and from 18% to 46%, respectively. However, this localized improvement was accompanied by a performance trade-off, resulting in a reduction in detection accuracy for other categories, such as Generic and Worms.

These findings highlight that, in the design of robust IDSs, it is highly recommended to combine multi-class classification models with specialized binary classifiers dedicated to those specific classes where the multi-class model faces severe identification challenges. Such a hybrid approach maintains an optimal balance between classification performance and computational overhead, which is critical for

environments with strict operational constraints, such as edge nodes or IoT gateways. This structural recommendation aligns with the multi-stage architecture proposed by [15], which effectively mitigates class imbalance and enhances the overall robustness of the detection system.

Consequently, future research will explore class-specific feature-selection mechanisms to determine the optimal sub-set of attributes for each attack family. Specifically, subsequent analyses will focus on how variations in individual features impact detection accuracy across different threat categories. The objective is to isolate the most discriminative attributes required for precise classification while strictly limiting the size of the feature set to mitigate the noise and redundancy inherent to high-dimensional data.

REFERENCES

[1] S. Sinha, "State of IoT 2025: Number of Connected IoT Devices Growing 14% to 21.1 Billion Globally," IoT Analytics, [Online], Available: https://iot-analytics.com/number-connected-iot-devices/, 2025.

[2] T. Fox, "Cybercrime to Cost the World $12.2 Trillion Annually by 2031," Cybersecurity Ventures, [Online], Available: https://cybersecurityventures.com/official-cybercrime-report-2025/, 2025.

[3] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Dataset for Network Intrusion Detection Systems (UNSW-NB15 Network Dataset)," Proc. of the IEEE 2015 Military Communications and Information Systems Conf. (MilCIS), pp. 1-6, Canberra, ACT, Australia 2015.

[4] S. More, M. Idrissi, H. Mahmoud and A. T. Asyhari, "Enhanced Intrusion Detection Systems Performance with UNSW-NB15 Data Analysis," Algorithms, vol. 17, no. 2, p. 64, 2024.

[5] M. Ahmad et al., "Intrusion Detection in Internet of Things Using Supervised Machine Learning Based on Application and Transport Layer Features Using UNSW-NB15 Data-set," EURASIP Journal on Wireless Communications and Networking, vol. 2021, no. 1, p. 10, 2021.

[6] M. Z. Hussain, A. Iftikhar, T. N. Usmani and M. Z. Hasan, "Leveraging Zero Trust Architecture for Network Intrusion Detection: A Comprehensive Evaluation Using the UNSW-NB15 Dataset," Spectrum of Engineering Sciences, vol. 3, no. 3, pp. 669-676, 2025.

[7] M. Jouhari, H. Benaddi and K. Ibrahimi, "Efficient Intrusion Detection: Combining $\chi 2$ Feature Selection with CNN-BiLSTM on the UNSW-NB15 Dataset," arXiv preprint, arXiv: 2407.14945, 2024.

[8] V. Sharma and M. Kumar, "Improving Intrusion Detection with Hybrid Deep Learning Models: A Study on CIC-IDS2017, UNSW-NB15 and KDD CUP 99," Journal of Information Systems Engineering and Management, vol. 10, no. 11, DOI: 10.52783/jisem.v10i11s.1665, 2025.

[9] A. D. Vibhute, M. Khan, C. H. Patil, S. V. Gaikwad, A. V. Mane and K. K. Patel, "Network Anomaly Detection and Performance Evaluation of Convolutional Neural Networks on UNSW-NB15 Dataset," Procedia Computer Science, vol. 235, pp. 2227-2236, 2024.

[10] M. Farhan et al., "Network-based Intrusion Detection Using Deep Learning Technique," Scientific Reports, vol. 15, no. 1, p. 25550, 2025.

[11] Z. Chkirbene, S. Eltanbouly, M. Bashendy, N. AlNaimi and A. Erbad, "Hybrid Machine Learning for Network Anomaly Intrusion Detection," Proc. of 2020 IEEE Int. Conf. on Informatics, IoT and Enabling Technologies (ICIoT), pp. 163-170, Doha, Qatar, 2020.

[12] M. H. Kabir et al., "Network Intrusion Detection Using UNSW-NB15 Dataset: Stacking Machine Learning Based Approach," Proc. of the 2022 IEEE Int. Conf. on Advancement in Electrical and Electronic Engineering (ICAEEE), pp. 1-6, Gazipur, Bangladesh, 2022.

[13] M. Belouch, S. El Hadaj and M. Idhammad, "Performance Evaluation of Intrusion Detection Based on Machine Learning Using Apache Spark," Procedia Computer Science, vol. 127, pp. 1-6, 2018.

[14] I. Mutambik, "An Efficient Flow-based Anomaly Detection System for Enhanced Security in IoT Networks," Sensors, vol. 24, no. 22, p. 7408, 2024.

[15] M. M. Mahmoud, Y. O. Youssef and A. A. Abdel-Hamid, "XI2s-IDS: An Explainable Intelligent 2-stage Intrusion Detection System," Future Internet, vol. 17, no. 1, p. 25, 2025.

ملخص البحث:

أدى انتشار البنى التحتية للشّبكات المترابطة وأجهزة إنترنت الأشياء إلى توسيع نطاق الهجمات الإلكترونية بشكلٍ كبير، مما يستدعي وجود أنظمة فعّالة للكشف عن الاختراقات تعتمد على التّعلُّم الآلي. وعلى الرغم من وجود مجموعات بيانات مرجعية، فإنّ خصائصها الرّسمية تفرض قيوداً فيما يتعلق بالمرونة وعدم توازن البيانات.

تعمل هذه الدّراسة على تقييم أثر تمثيل بيانات مخصَّص من خلال إنشاء مجموعة بيانات جديدة من ملفّات PCAP. وقد طبّقنا آلية عملٍ لتصنيف الحزم وتجميع التّدفقات أحادية الاتّجاه، واستخراج مجموعةٍ مصغّرة من 21 ميزة، وقارنا هذا التّمثيل مع مجموعة البيانات الرّسمية الّتي تضمّ 49 ميزة باستخدام بنى تعلّم آلي مختلفة في مهامّ التّصنيف الثّنائي ومتعدّد الفئات.

وتشير النّتائج إلى أنّ مجموعة البيانات المخصّصة تحقّق أداءً تنافسياً على الرّغم من الانخفاض الكبير في حجم الملفّ وعدد الميزات. والجدير بالذّكر أنّ التّمثيل المخصّص يوازن بفاعلية بين دقّة الكشف والكفاءة الحسابية، ممّا يوفّر استراتيجية فعّالة للبيئات ذات القيود التّشغيلية الصّارمة، مثل عُقَد الحافة او بوابات إنترنت الأشياء.

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

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

ISSN 2415 - 1076 (Online)
ISSN 2413 - 9351 (Print)

حزيران ٢٠٢٦ المجلد ١٢ العدد ٢

JJCIT

الصفحات	عنوان البحث

www.jjcit.org

jjcit@psut.edu.jo

مجلة علمية عالمية متخصصة تصدر
بدعم من صندوق دعم البحث العلمي والابتكار