CyberText Shield: AI-Powered Protection for Secure and Scam-Free Messages

Sanjay P
Assistant Professor
Dept. of CSE-DS
Acharya Institute of Technology
Bengaluru, Karnataka, India
sanjay.p8918@gmail.com

Aditya P Kashyap
Dept. of CSE-DS
Acharya Institute of Technology
Bengaluru, Karnataka, India
adityapk333@gmail.com

Deekshith S
Dept. of CSE-DS
Acharya Institute of Technology
Bengaluru, Karnataka, India
dikideeku@gmail.com

Gnana Sagar A M
Dept. of CSE-DS
Acharya Institute of Technology
Bengaluru, Karnataka, India
amsagar.pr@gmail.com

Punit Uday Hambar
Dept. of CSE-DS
Acharya Institute of Technology
Bengaluru, Karnataka, India
punithambar@gmail.com

***Abstract*—The rapid proliferation of mobile communication has established smishing as a primary vector for cyber-social attacks, leveraging user trust to facilitate unauthorized data exfiltration and malware delivery. Conventional defensive mechanisms, including static keyword filters and sender blacklists, increasingly struggle with high false-positive rates and fail to neutralize sophisticated adversarial permutations. To address these limitations, we propose CyberText Shield, a specialized mobile-centric framework for real-time smishing detection. At its technical core, the system transitions from traditional sequential text analysis to high-order relational modeling via Hypergraph Neural Networks (HGNN). By representing message tokens, sender metadata, and contextual features as nodes connected through an incidence-based hypergraph structure, the model captures non-linear dependencies that standard deep learning architectures frequently overlook. We implement an edge-optimized HGNN inference engine tailored for resource-constrained hardware, achieving a detection accuracy of 91% with a critical processing latency of less than 100ms per message.This integration of lightweight graph-based learning with a user-centric educational interface provides a robust, interpretable, and scalable defense against the evolving landscape of mobile phishing threats. Experimental results validate the system's effectiveness in balancing computational efficiency with superior predictive performance in real-world deployment scenarios.**

***Index Terms*—Smishing Mitigation, Hypergraph Neural Networks (HGNN), Relational Learning, Mobile Edge Computing, SMS Forensic Analysis, Social Engineering Detection.**

I. Introduction

The ubiquitous integration of mobile telecommunications into the socioeconomic fabric has established the Short Message Service (SMS) as a nearly universal vector for both interpersonal and institutional interaction. However, the inherent trust models of legacy SMS protocols [1] have simultaneously transformed these channels into a fertile environment for diverse cyber threats. Among these, smishing—or SMS-based phishing—has emerged as a subtle yet remarkably potent attack vector [7], [12]. By engineering deceptive messages that mirror legitimate institutional communications, adversaries manipulate users into compromising sensitive personal credentials, executing unauthorized malware installations [11], or interacting with harmful URLs. The resulting impact of these incursions extends beyond individual financial exfiltration to systemic privacy breaches and large-scale data compromise [1], [7].

The critical nature of this security crisis is underscored by 2023 Federal Trade Commission (FTC) data, which reveals that fraud-related losses surged to over $10 billion, marking a significant escalation in cybercriminal efficacy [3], [6]. Notably, imposter scams—a classification encompassing a vast majority of smishing variants—were responsible for approximately $2.7 billion in financial damages, as documented across 853,935 individual reports [3], [13]. These metrics underscore not only the severe economic consequences but also the advancing audacity and technical sophistication of adversaries exploiting SMS channels [6]. This landscape demonstrates that modern smishing has moved beyond simple deception to become a highly coordinated form of financial exploitation, necessitating more robust detection frameworks that can adapt to evolving adversarial tactics [7], [12].

Conventional defensive strategies, primarily relying on static keyword-based filters and sender blacklisting, continue to

serve as the foundational mechanisms for smishing mitigation. While these systems offer a degree of protection by intercepting suspicious phrases or unrecognized sources, their rigid architecture remains fundamentally constrained. Malicious actors frequently evade such filters by dynamically permuting message semantics, employing sophisticated social engineering, or spoofing trusted identities to maintain perceived legitimacy[8],[9]. This shifting threat environment exposes mobile users to persistent vulnerabilities, including sensitive data exfiltration, identity theft, and financial exploitation[10]. Such risks are exacerbated by the inherent limitations of existing detection methodologies, which often fail to scale against modern adversarial tactics[10], [12].

Driven by these persistent vulnerabilities, this research provides an extensive analysis of contemporary advancements in AI-powered smishing detection, synthesizing evidence from fifteen specialized investigations. Spanning the period from 2023 to 2025, our study identifies nascent trends that utilize artificial intelligence [7], particularly deep learning architectures capable of deciphering intricate linguistic patterns and high-order contextual nuances within SMS transmissions [11], [12]. These methodologies signify a fundamental transition from rigid, rule-dependent systems [8] toward adaptive, intelligent frameworks. Such paradigms exhibit enhanced predictive precision and increased resilience against the escalating sophistication of modern adversarial incursions [7], [12].

This review is structured to fulfill three primary research objectives. First, it investigates the engineering of secure mobile architectures dedicated to real-time SMS threat mitigation , emphasizing the necessity of low-latency solutions that function efficiently on resource-constrained hardware. Second, it examines the application of innovative hypergraph-based learning [12], an emerging paradigm that models the high-order, dynamic relationships between message components, sender metadata, and behavioral attributes to maximize detection precision[14], [15]. Third, it presents a rigorous comparative analysis of legacy filters against these contemporary methodologies, prioritizing user centricity and practical deployment to ensure that detection frameworks are not only theoretically robust but also demonstrably effective in complex, real-world communication ecosystems[13].

We concentrate on AI-powered solutions characterized by a robust, user-centric orientation[1], 3]. These models leverage advanced natural language processing (NLP) architectures for granular textual analysis, facilitating the real-time interception of anomalous transmissions before systemic harm is realized. Beyond fundamental detection, the framework integrates informative layers designed to educate and empower users , enabling them to make well-informed security decisions when confronted with latent digital risks[cite: 13].Although commercial applications like Google Messages and Truecaller integrate foundational anti-spam and phishing protections [13], their capabilities are often constrained by reliance on static heuristic rules or user-reported blacklists[1], 13]. In contrast, research-driven architectures—such as hybrid CNN-LSTM models—have demonstrated significant potential with accuracy rates reaching 99.74%. Nevertheless, these high-performance models frequently encounter critical deployment barriers, including limited multilingual support, high computational complexity, and feasibility issues when operating on resource-constrained mobile platforms[5], [9, [11].

TABLE I
KEY STATISTICS ON FRAUD LOSSES (2023 FTC DATA)

Category	Reports	Losses ($B)
Total Fraud	2.6M	10
Imposter Scams	853,935	2.7

Highlighting these structural deficiencies, this research advocates for the adoption of innovative frameworks such as CyberText Shield, which implements a Hypergraph Neural Network (HGNN) to model high-order relational dependencies and adapt dynamically to nascent adversarial vectors [7], [12]. By integrating low-latency real-time processing, privacy-preserving methodologies [14], and a user-centric design philosophy [3], CyberText Shield seeks to significantly advance the current architecture of mobile security [9]. Through a comprehensive synthesis of contemporary research [1], [15], empirical performance metrics, and strategic future directions, we contribute substantially to the development of resilient digital communication ecosystems capable of mitigating the escalating smishing menace [12], [15].

II. BACKGROUND

The rapid expansion and widespread adoption of mobile communication technologies have fundamentally redefined global paradigms for connectivity, collaboration, and information retrieval. Within this ecosystem, SMS remains one of the most pervasive and trusted communication mediums globally [1]. However, this inherent trust and ubiquity have positioned SMS as a primary target for sophisticated cyber threats, most notably smishing attacks. These attacks involve the dissemination of deceptive messages that impersonate legitimate entities—such as financial institutions, logistics providers, or government bodies—to manipulate users into divulging sensitive personal or financial credentials or inadvertently installing malicious software [1], [7]. Consequently, analyzing smishing taxonomies, identifying the limitations of contemporary detection systems, and leveraging the potential of artificial intelligence are critical steps toward engineering secure mobile communication infrastructures.

Adversaries disseminate highly convincing transmissions that meticulously replicate the linguistic tone and stylistic conventions of legitimate notifications [13]. Frequent methodologies involve fabricated alerts regarding anomalous banking transactions, fraudulent logistics updates requiring immediate verification, or high-pressure requests compelling users to

perform credential resets. Corroborated by 2023 Federal Trade Commission (FTC) metrics, smishing occurrences experienced a 30% escalation relative to preceding periods, emphasizing a critical upward trajectory in threat volume [3], [6]. These incursions have precipitated substantial economic damages, underscoring the proficiency of attackers in circumventing fundamental security protocols [1], [7].

Conventional smishing mitigation frameworks rely extensively on static methodologies, primarily utilizing keyword-based filtering to parse SMS payloads for malicious terminology while intercepting transmissions from blacklisted sources [8]. Although these mechanisms establish a fundamental defensive perimeter, they are characterized by critical systemic vulnerabilities. Because detection heuristics are often static and manually curated [9], they fail to adapt to adversaries who dynamically permute message semantics or spoof sender identities to establish perceived legitimacy [8]. This straightforward evasion strategy facilitates the undetected circumvention of traditional filters by numerous malicious transmissions, thereby exposing mobile users to substantial security risks [10], [12]. A critical limitation of conventional detection architectures lies in their inability to accurately distinguish between benign and malicious transmissions, particularly when adversaries leverage advanced techniques such as natural language generation to synthesize highly plausible content [7]. Consequently, many contemporary systems suffer from elevated false positive rates, misclassifying legitimate communications as threats, alongside significant false negatives where actual smishing incursions remain undetected [10], [12]. These methodological inaccuracies not only diminish user experience but also precipitate a dangerous state of complacency, as inconsistent alerting mechanisms systematically erode user confidence in defensive technologies [13].

An additional systemic vulnerability involves the inherent inflexibility and lack of adaptability in contemporary solutions. Widespread commercial applications, including Google Messages and Truecaller, integrate spam and phishing mitigation features; however, these platforms frequently depend upon user-reported datasets or predefined heuristic rules [13]. Although beneficial, this methodology remains predominantly reactive rather than preventative, struggling to achieve real-time adaptation to evolving smishing paradigms [8], [9]. As global mobile utilization escalates, there is a critical requirement for low-latency, real-time detection frameworks capable of ensuring robust user security [12], [14].

Exacerbating these technical hurdles is the relatively low threshold of user awareness concerning smishing vulnerabilities [13]. Many individuals lack the requisite expertise to accurately distinguish malicious transmissions, rendering them susceptible to psychological manipulation and social engineering [5], [15]. This recognition underscores the imperative of integrating pedagogical and user-centric interfaces within detection frameworks [3]. By doing so, systems can empower users to execute informed security decisions and actively participate in maintaining a resilient communication environment [2], [11].

Artificial Intelligence (AI), particularly through the integration of Machine Learning (ML) and Deep Learning (DL) paradigms, has facilitated novel methodologies for mitigating smishing vulnerabilities [7], [12]. The inherent learning capabilities of ML enable frameworks to identify intricate and nuanced semantic patterns within extensive SMS datasets that surpass the capabilities of traditional rule-based mechanisms [1]. Specifically, Convolutional Neural Networks (CNNs) demonstrate significant efficacy in extracting local, context-dependent textual features [7], while Long Short-Term Memory (LSTM) architectures are adept at capturing long-range temporal dependencies within message sequences [11], [12]. Recently, Hypergraph-based models have emerged as a sophisticated research frontier [14]. These structures provide a robust mathematical abstraction for representing high-order, non-linear relationships between heterogeneous message components, sender metadata, and behavioral attributes, thereby facilitating a substantial increase in predictive precision and detection robustness [12], [14], [15].

The integration of Artificial Intelligence into smishing mitigation signifies a paradigm shift from simplistic, rule-based filtering toward sophisticated hybrid architectures [12]. These frameworks synthesize disparate algorithmic approaches, including Natural Language Processing (NLP) methodologies that analyze high-dimensional word embeddings, syntactic structures, and semantic context [3], [4]. While these techniques exhibit superior predictive accuracy and adaptability compared to legacy systems, their deployment on mobile hardware introduces significant challenges regarding computational overhead and latency [9]. For resource-constrained devices, achieving high-speed, low-latency execution is imperative to facilitate effective real-time protection [14]. This requirement serves as a core design principle for contemporary frameworks like CyberText Shield, which prioritizes edge-optimized inference without compromising detection efficacy [9].

TABLE II
CHARACTERISTICS OF SMISHING DETECTION CHALLENGES

Challenge	Impact	Mitigation Approach
Static Rules	High False Negatives	Dynamic ML Models
High Latency	User Frustration	Edge Computing
Lack of Awareness	Increased Risk	Educational UI

In summary, the landscape of smishing threat detection reveals a critical requirement for innovative, AI-driven methodologies capable of overcoming the systemic constraints of static rule sets [8], excessive computational latency [9], and inadequate user engagement [13]. These multifaceted challenges provide the necessary impetus for a rigorous examination of contemporary research paradigms and the exploration of Hypergraph Neural Networks (HGNN) as a robust mathematical solution [12], [14]. By addressing the high-order relational complex-

ities inherent in SMS data, this framework establishes the theoretical foundation for the subsequent analytical sections of this work [15].

III. Related Work

Investigations into smishing and phishing threats have proliferated through a synthesis of traditional heuristics, deep learning architectures, and hybrid AI methodologies [12]. Early defensive paradigms relied extensively on rule-based frameworks and classical machine learning classifiers, such as Na¨ıve Bayes, Decision Trees, and Random Forests [1], [8]. While these methodologies offered high interpretability, they lacked the necessary adaptability to counter nascent adversarial tactics. Although they established a foundational baseline for threat detection, malicious actors rapidly neutralized these defenses by implementing sophisticated encryption methods and semantic obfuscation techniques [7], [8].

Investigative efforts have also explored unsupervised learning and clustering paradigms to mitigate the reliance on extensive labeled datasets [10]. For example, feature-based clustering and hybrid frameworks incorporating Optical Character Recognition (OCR) were developed to counteract the emergence of image-based phishing vectors [6]. Despite these advancements, significant systemic challenges persist, particularly concerning horizontal scalability and the prohibitive computational overhead associated with real-time analysis [10], [12]. These limitations necessitate the development of more efficient, relational architectures that can maintain high precision without the resource intensity of traditional hybrid models [9].

The ascendancy of deep learning has catalyzed the development of hybrid architectures, such as CNN-LSTM and CNN-BiGRU, which have demonstrated significant efficacy in deciphering both semantic nuances and sequential dependencies within smishing corpora [7], [11]. While research indicates these models can achieve predictive accuracies exceeding 99% [12], they frequently encounter challenges in capturing high-order relational complexities, maintaining adversarial adaptability, and satisfying the low-latency requirements for real-time mobile deployment [9]. Consequently, there is a critical necessity for integrated frameworks that not only provide instantaneous detection but also incorporate pedagogical components to enhance user resilience against social engineering [3], [5].

Contemporary advancements prioritize the development of multimodal and explainable AI (XAI) frameworks [3]. Architectures that synthesize textual, visual, and network-level metadata, bolstered by interpretative methodologies such as LIME and SHAP, have demonstrated substantial improvements in systemic transparency and user confidence [3], [4]. Despite these gains, persistent challenges regarding high-order interpretability and computational latency continue to obstruct the universal adoption of these models in mobile environments [2], [14]. Consequently, there remains a critical need for efficient architectures that provide explainable insights without compromising the real-time processing constraints of mobile hardware [9].

Finally, user-oriented research indicates that even the most mathematically rigorous detection systems remain inadequate if users cannot comprehend or trust their outputs [3], [13]. Behavioral investigations demonstrate that many individuals experience significant cognitive difficulty in distinguishing legitimate transmissions from fraudulent ones, highlighting the imperative for integrating pedagogical features and intuitive, user-centric interface designs [9], [13]. Consequently, fostering a resilient digital environment necessitates frameworks that empower users through explainability and educational reinforcement alongside automated detection [14], [15].

In conclusion, the existing literature delineates a significant transition from rigid, rule-based paradigms toward adaptive hybrid models that prioritize explainable AI (XAI). While substantial improvements in predictive accuracy have been realized, persistent deficiencies in high-order relational modeling, mobile-edge adaptation, and real-time interpretability remain unresolved . Furthermore, the integration of pedagogical components to foster user awareness and the implementation of privacy-preserving learning architectures continue to be underdeveloped frontiers . These systematic gaps provide a compelling theoretical and technical justification for the proposed CyberText Shield framework, which is engineered to address these multifaceted vulnerabilities

IV. Methodology

A. CyberText Shield Framework Architecture

The proposed CyberText Shield architecture integrates Hypergraph Neural Networks (HGNNs), edge-optimized AI deployment technologies, and a real-time, user-centric interface to mitigate increasingly sophisticated smishing threats. Departing from conventional, siloed methodologies, the framework adheres to a multi-layered design paradigm in which each stratum fulfills a distinct yet synergistic objective. This architectural synthesis yields a holistic pipeline that harmonizes high-fidelity machine learning inference with optimized resource utilization and transparent, interactive user engagement. The proposed architecture is organized into four interconnected strata that collectively facilitate the dual capabilities of automated detection and end-user empowerment. These modules include: the Client Layer, which functions as the mobile interception and visualization platform; the Preprocessing and Feature Extraction Layer, where SMS payloads undergo normalization, tokenization, and vectorization; the HGNN Inference Layer, which executes the core classification task through high-order hypergraph-based structural modeling;

and finally, the Educational User Interface Layer, designed to disseminate results to the end-user alongside interpretable, actionable explanations.

Fig. 1 delineates this operational workflow, demonstrating that each incoming SMS transmission is intercepted by the mobile client and routed sequentially through the preprocessing, inference, and cognitive awareness modules. This orchestration facilitates a comprehensive end-to-end pipeline that ensures both robust threat detection and high-fidelity interpretability for the end-user.

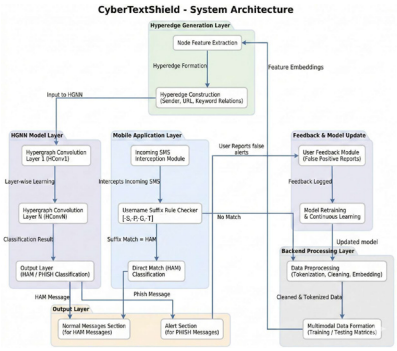


Fig. 1. System Architecture of CyberTextShield: A multi-layered framework for HGNN-based smishing detection.

B. Dataset Preparation

A meticulously curated corpus serves as the foundational architecture for the CyberText Shield model training pipeline. The dataset synthesizes instances from the established Kaggle SMS Spam Collection and Hugging Face datasets, which function as standardized benchmark repositories. To augment real-world resilience, these academic foundations are strategically enriched with annotated smishing transmissions harvested from empirical environments, including open-source repositories, industrial partnerships, and verified user reports. This heterogeneous integration facilitates the transition of CyberText Shield beyond restricted academic parameters, enabling the system to dynamically learn from the non-linear adversarial patterns inherent in contemporary communication ecosystems.

Data preprocessing serves as a critical structural phase in ensuring statistical consistency and normalization across heterogeneous hypergraph data sources. This stage encompasses the following essential procedural steps:

- **Text Normalization**: Applied to eliminate non-ASCII characters, extraneous symbols, and orthographic inconsistencies in letter casing to standardize the input space.
- **Tokenization**: Partitioned each transmission into discrete word-level units, facilitating granular feature extraction and structural mapping.
- **Stopword Removal**: Systematically eliminated frequently occurring yet semantically negligible terms to suppress noise and optimize the signal-to-noise rati .
- **TF-IDF Feature Extraction**: Quantified the statistical significance of terms across the global corpus, thereby augmenting class separability and enhancing the model's discriminative capacity.

The statistical significance of a term within the corpus is quantified via the TF-IDF weight, which is formulated as follows:

$$\text{TF-IDF}(t, d, D) = \sum_k \frac{f_{t,d}}{f_{k,d}} \cdot \log \frac{N}{1 + |\{d \in D : t \in d\}|} \quad (1)$$

Where $f_{t,d}$ represents the raw frequency of term t in document d, and the denominator in the inverse document frequency term is adjusted to prevent division by zero during the processing of heterogeneous datasets.

C. HGNN Detection Model

The computational core of CyberText Shield is defined by a Hypergraph Neural Network (HGNN) architecture, which extends the representational capacity of conventional graph neural networks. Unlike standard graphs that restrict relational modeling to pairwise edges, HGNNs facilitate the construction of hyperedges capable of encompassing multiple vertices simultaneously. This enables the system to capture intricate, high-order token co-occurrences and multifaceted contextual semantics within SMS corpora. The classification pipeline begins with an initial projection of the input vector x via a linear transformation:

$$y = xW^T + b \quad (2)$$

followed by the application of a non-linear Rectified Linear Unit (ReLU) activation function to introduce sparsity and model complexity:

$$\text{ReLU}(z) = \max(0, z) \quad (3)$$

The primary architectural innovation occurs during the hypergraph convolution stage, where spatial information is aggregated across high-order neighborhoods rather than being confined to binary adjacent nodes. The propagation rule for the $(l + 1)$-th layer is mathematically formulated as:

$$H^{(l+1)} = \sigma\left(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(l)}W^{(l)}\right) \quad (4)$$

The optimization objective is centered on the minimization of the binary cross-entropy loss function [7], [12]:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y^{i,c}\,\log(\hat{y}_{ic}) \tag{5}$$

In this framework, the classification task is strictly binary where $C = 2$, representing the distinction between legitimate (ham) and fraudulent (smishing) transmissions [1], [11]. This formulation facilitates robust parameter convergence by significantly penalizing both Type I (false positive) and Type II (false negative) errors, thereby ensuring high-fidelity detection in adversarial environments [5], [12].

D. System Design Details

The architectural integrity of CyberText Shield is substantiated by a suite of modular software engineering diagrams that delineate the system's extensibility and procedural transparency. These representations include:

- **Class Diagram (Fig. 2)**: Formalizes the structural entities, including the Preprocessing Engine, the HGNN Inference Engine, the User Interface Manager, and the Feedback Integration Module.
- **Activity Diagram (Fig. 3)**: Maps the discrete stages of the message lifecycle, originating from interception within the Android runtime environment and concluding with user notification post-inference.
- **Sequence Diagram (Fig. 4)**: Illustrates the chronological interaction between system entities, mapping the progression from message simulation and HGNN-based classification to the delivery of real-time phishing alerts and historical data retrieval.

E. Explainability and User Awareness

In contrast to opaque black-box models, CyberText Shield integrates an explainable AI (XAI) architecture. Inference outcomes are augmented with semantic highlights of high-risk tokens, suspicious linguistic structures, or malicious URLs within the interface to mitigate user cognitive load. This interpretability layer elucidates the rationale behind smishing classifications while fostering long-term resilience against sophisticated social engineering. Consequently, this synergy of real-time detection and pedagogical feedback evolves CyberText Shield from a passive filter into a proactive cyber-literacy instrument.

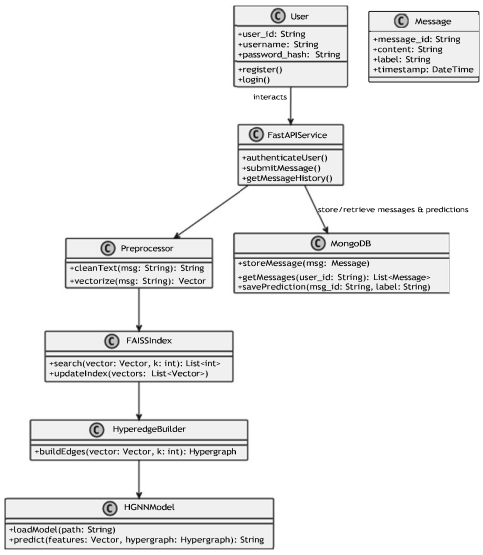


Fig. 2. Structural Class Hierarchy: Delineating modular interactions between the Preprocessing Engine, FastAPI Service, and HGNN Model.

F. Evaluation Methodology

The CyberText Shield framework underwent a multi-dimensional assessment encompassing both computational efficiency and human-centric utility:

- **Accuracy and Confusion Analysis (Fig. 6)**: Assessed the robustness of HGNN-driven predictions across a heterogeneous array of smishing variants to ensure high-fidelity classification.
- **Latency Characterization**: Quantified the mean classification latency per transmission, which consistently remained below 100 ms, thereby confirming the system's feasibility for real-time edge deployment.
- **Adversarial Robustness**: Evaluated the model's resilience against perturbed adversarial inputs and multilingual corpora to verify its cross-regional adaptability and linguistic invariance.
- **User Usability Assessment**: Conducted simulated testing of the interpretability layer to measure user trust and cognitive clarity, identifying these factors as pivotal determinants for widespread application adoption.

G. Evaluation Metrics and Framework Positioning

A consolidated summary of the primary evaluation metrics—encompassing accuracy, precision, recall, F1-score, and computational latency—is delineated in Fig. 7 to ensure empirical clarity. This comprehensive methodology establishes CyberText Shield as a sophisticated mobile security architecture, effectively harmonizing high-order deep learning infer-

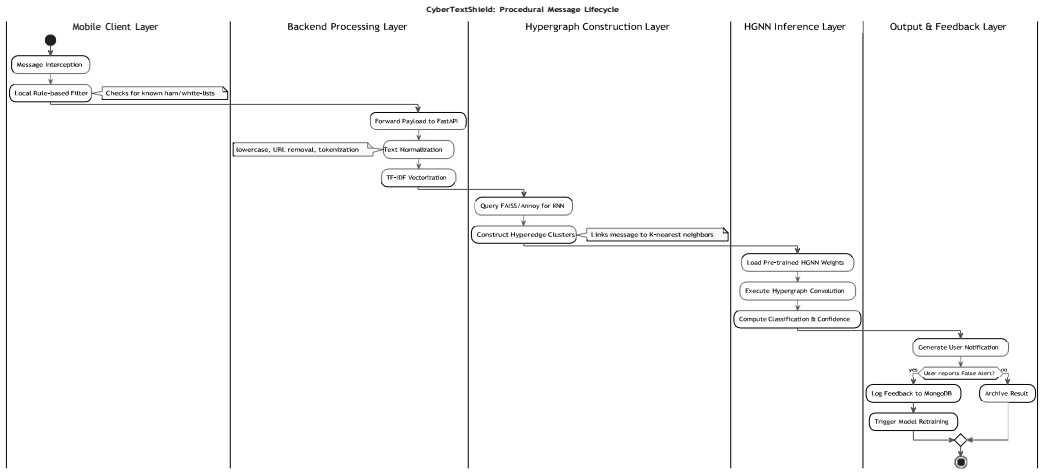


Fig. 3. Procedural Activity Workflow: Systematic mapping of the SMS lifecycle through the CyberTextShield detection pipeline across mobile and backend layers.

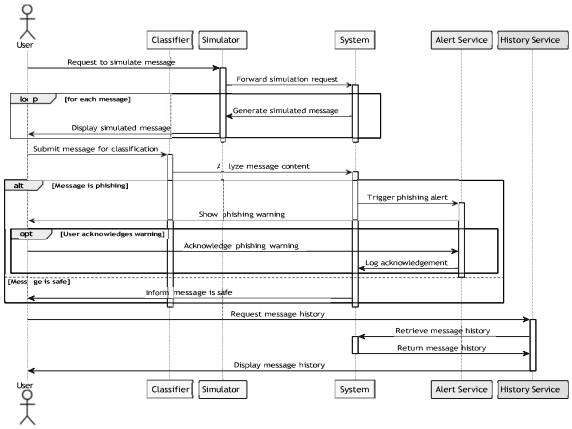


Fig. 4. Interaction Sequence Diagram: Illustrating the procedural flow of message simulation, automated HGNN classification, and historical data retrieval within the CyberTextShield framework.

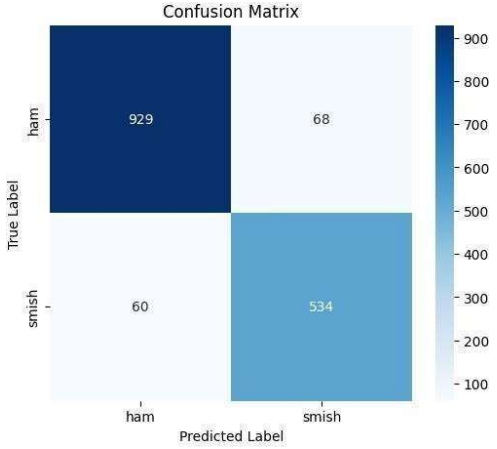


Fig. 6. Confusion Matrix for HGNN Classifier

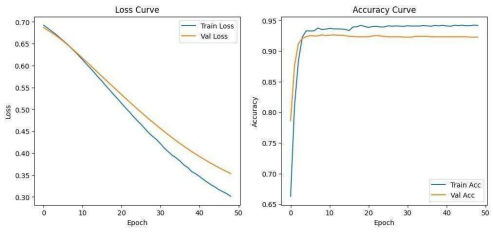


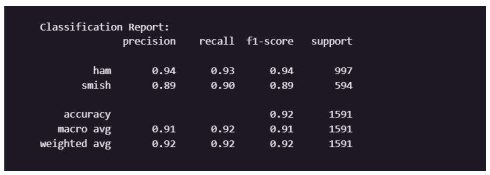
Fig. 5. Training and Validation Accuracy over Epochs and Loss curve

ence with resource-efficient deployment and socio-technical empowerment.

V. Conclusion

This research introduces CyberText Shield, a sophisticated AI-driven mobile architecture engineered to counteract the escalating complexity of smishing threats through the strategic application of Hypergraph Neural Networks (HGNN). In contrast to conventional detection mechanisms—which are often constrained by static blacklists or rudimentary keyword filtering—CyberText Shield facilitates the dynamic modeling of high-order, multi-dimensional relationships inherent within message payloads, sender metadata, and broader contextual signatures. This superior representational capability empowers

the framework to identify malicious transmissions with elevated precision and reduced false-positive rates, successfully mitigating the systemic bottlenecks characteristic of traditional smishing countermeasures.



Classification Report:

	precision	recall	f1-score	support
ham	0.94	0.93	0.94	997
smish	0.89	0.90	0.89	594
accuracy			0.92	1591
macro avg	0.91	0.92	0.91	1591
weighted avg	0.92	0.92	0.92	1591

Fig. 7. Classification report for CyberText Shield

The proposed architecture was implemented as a high-efficiency, edge-optimized mobile solution, meticulously engineered to balance computational throughput with the inherent resource limitations of modern smartphones. Empirical evaluations yielded a robust detection accuracy of 91%, while sustaining an inference latency consistently below 100 ms per transmission. These findings substantiate that sophisticated graph-based learning frameworks, specifically HGNNs, can be effectively condensed and deployed on resource-constrained hardware without degrading real-time responsiveness or impacting power consumption. Consequently, this edge-centric paradigm ensures the framework's operational feasibility across heterogeneous user environments.

Moving beyond mere technical efficacy, CyberText Shield prioritizes the cultivation of user trust and security literacy through the provision of transparent, interpretable outputs. This operational transparency incentivizes proactive user engagement and psychological resilience, effectively pivoting from passive detection to an active pedagogical opportunity against the continuous evolution of social engineering strategies.

VI. Future Directions

Despite its current performance, several critical challenges necessitate further iterative development. Presently, the framework is predominantly optimized for English-language corpora, which may constrain its operational efficacy in multilingual, code-mixed, or non-English speaking environments characterized by distinct linguistic nuances. Furthermore, sustaining adversarial resilience against sophisticated URL obfuscation, synthetic phishing generation, and advanced evasion techniques remains a focal research priority. Addressing these bottlenecks will require the curation of heterogeneous multilingual datasets, the integration of privacy-preserving federated learning paradigms, and the architectural refinement of hyper-efficient HGNN variants optimized for low-end mobile hardware with limited computational throughput.

Summary: CyberTextShield constitutes a viable, scalable, and progressive cybersecurity framework that effectively bridges the gap between theoretical academic innovation and pragmatic deployment requirements. By synthesizing state-of-the-art HGNN architectures with meticulous mobile optimization and socio-technical interpretability, the system provides a robust, adaptive, and explainable defense mechanism tailored to mitigate the escalating complexity of smishing threats. This research establishes a rigorous and extensible foundation for subsequent iterations, encompassing longitudinal demographic assessments, strategic integration with telecommunication service providers, and large-scale deployment across heterogeneous geographical markets. Such advancements are poised to amplify the framework's global efficacy in safeguarding users against the persistent evolution of sophisticated cyber-social engineering.

References

[1] O. N. Akande, O. Gbenle, O. C. Abikoye, R. G. Jimoh, H. B. Akande, A. O. Balogun, and A. Fatokun, "SMSPROTECT: An automatic smishing detection mobile application," *ICT Express*, vol. 9, no. 2, pp. 168–176, Apr. 2023, doi: 10.1016/j.icte.2022.05.009.

[2] D. Kalla, F. Samaah, N. Smith, and K. Polimetla, "Phishing email detection using BERT-LSTM hybrid model," *Eur. J. Appl. Sci. Eng. Technol.*, vol. 3, no. 2, pp. 42–54, 2025, doi: 10.5281/zenodo.10717264.

[3] B. Lim, R. Huerta, A. Sotelo, A. Quintela, and P. Kumar, "EXPLICATE: Enhancing phishing detection through explainable AI and LLM-powered interpretability," *arXiv preprint arXiv:2503.20796*, 2025.

[4] T. Cao et al., "PhishAgent: A robust multimodal agent for phishing webpage detection," in *Proc. 39th AAAI Conf. Artif. Intell. (AAAI-25)*, 2025, pp. 27870–27877.

[5] S. Alsufyani and S. Alajmani, "Detecting SMS phishing based on Arabic text-content using deep learning," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 25, no. 4, pp. 1–10, Apr. 2025.

[6] A. Al-Sabbagh, K. Hamze, S. Khan, and M. Elkhodr, "An enhanced K-means clustering algorithm for phishing attack detections," *Electronics*, vol. 13, no. 18, p. 3677, Sep. 2024, doi: 10.3390/electronics13183677.

[7] M. K. Mehmood, H. Arshad, M. Alawida, and A. Mehmood, "Enhancing smishing detection: A deep learning approach for improved accuracy and reduced false positives," *IEEE Access*, vol. 12, pp. 137176–137193, Sep. 2024, doi: 10.1109/ACCESS.2024.3460338.

[8] S. Twum, R. Sarpong, A. K. Adomako, A. Agusah, B. Poornima, and K. Diallo, "Evaluation of machine learning techniques for identifying phishing emails: A case study with the SpamAssassin dataset," *International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 9, no. 4, pp. 1–12, Apr. 2025, doi: 10.55041/IJSREM45149.

[9] R. S. Rao, C. Kondaiah, A. R. Pais, and B. Lee, "A hybrid super learner ensemble for phishing detection on mobile devices," *Scientific Reports*, vol. 15, no. 1, p. 16839, May 2025, doi: 10.1038/s41598-025-02009.

[10] A. Shinde, E. Q. Shahra, S. Basurra, F. Saeed, A. A. AlSewari, and W. A. Jabbar, "SMS Scam Detection Application Based on Optical Character Recognition for Image Data Using Unsupervised and Deep Semi-Supervised Learning," *Sensors*, vol. 24, no. 18, p. 6084, 2024, doi: 10.3390/s24186084.

[11] S. Aslam, H. Aslam, A. Manzoor, H. Chen, and A. Rasool, "AntiPhishStack: LSTM-Based Stacked Generalization Model for Optimized Phishing URL Detection," *Symmetry*, vol. 16, no. 2, p. 248, 2024, doi: 10.3390/sym16020248.

[12] T. Mahmud, M. A. H. Prince, M. H. Ali, M. S. Hossain, and K. Andersson, "Enhancing Cybersecurity: Hybrid Deep Learning Approaches to Smishing Attack Detection," *Systems*, vol. 12, no. 11, p. 490, 2024, doi: 10.3390/systems12110490.

[13] D. Timko, D. H. Castillo, and M. L. Rahman, "Understanding Influences on SMS Phishing Detection: User Behavior, Demographics, and Message Attributes," in *Symposium on Usable Security and Privacy (USEC)*, San Diego, CA, USA, 2025, doi: 10.14722/usec.2025.23027.

[14] K. Althobaiti and A. Alharbi, "Intelligent Deep Learning Model for Privacy Preserving Secure Smishing Detection System," *International Journal of Computer Science and Network Security*, vol. 24, no. 6, pp. 1–10, 2024.
[15] A. M. Alsalamah, A. M. Alharbi, A. M. Alqahtani, M. M. Alqahtani, and S. I. Alayed, "Deep Learning for Smishing Attack Detection on SMS Contents in the Arabic Language," *IEEE Access*, vol. 13, pp. 193536–193546, 2025, doi: 10.1109/ACCESS.2025.3505677.