Grenze International Journal of Engineering and Technology, June Issue

Explainable AI-based Phishing Detection using TF-IDF-SVD Embeddings and XGBoost Classification

Jane Rubel Angelina Jeyaraj[1], B. Vamsi Krishna[2], B. Narendra Reddy[3], N. Harshal Karthik[4], N. Harsha Vardhan Reddy[5] and Vadlamudi Venkata Naveen[6]
[1-5]Kalasalingam Academy of Research and Education/Department of Computer Science and Education, Krishnankoil, India
Email: j.janerubelangelina@klu.ac.in, vamsikrishnabajjuri3@gmail.com, narendrareddybadam553@gmail.com, exhaniko64@gmail.com, nagulapatiharsha@gmail.com
[6]College of Engineering Tamkang University, Taiwan, Tamsui, New Taipei City, Taiwan
Email: venkatanaveenvadlamudi@gmail.com

***Abstract*— The AI-powered phishing detection system in this work, PhishSentinel, is designed to enhance cybersecurity by accurately identifying malicious emails and URLs. In this approach, a hybrid machine learning model is developed fusing semantic analysis using TF-IDF and SVD embeddings of text with handcrafted URL and text features; an XGBoost model conducts classification for precise detection. A web application based on Streamlit allows real-time analysis, model retraining, and visual explanation of predictions as SHAP-style feature importance charts. This approach improves not only the accuracy of detection but also interpretability to help users understand why an email or URL has been flagged. Contributing to safer digital environments, PhishSentinel directly supports SDG 9: Industry, Innovation, and Infrastructure by promoting secure and resilient digital infrastructure using innovative AI solutions. The proposed approach provides a scalable, explainable, and easy-to-use solution for mitigating phishing attacks in modern communication systems.**

***Index Terms*— Phishing detection, AI security, TF-IDF, SVD, XGBoost, Streamlit, explainable AI, SDG 9: Industry, Innovation and Infrastructure, semantic analysis, feature engineering.**

I. INTRODUCTION

Phishing remains one of the most frequent and dangerous cyber threats, attacking individuals, organizations, and digital infrastructures worldwide. While online communications grow, malicious e-mail and URL detection has been more challenging due to emerging attack patterns and advanced social engineering methods. Within this scenario, PhishSentinel will be an intelligent AI-powered system tailored to enhance digital security through fast, reliable, and explainable phishing detection. The overall objective is thus to develop a robust and accessible platform for empowering users toward threat identification before the latter cause damage.

It focuses on what truly counts in phishing detection: making sense of the content, structure, and intent of suspicious messages. PhishSentinel will go through not just the visible text but also hidden URL patterns, domain characteristics, linguistic features, and phishing-related keywords. This combination of semantic meaning and structural analysis provides the ability to identify not just the traditional phishing attempts but the advanced, carefully crafted ones that often slip past rule-based filters.

Grenze ID: 01.GIJET.12.2.531_1

The need for such a system arises from the ever-increasing scale and complexity of cyberattacks, which call for more intelligent methods beyond simple signature matching. Indeed, PhishSentinel improves over time by learning with either genuine or malicious examples using various machine learning models. It is thus suitable for modern digital environments where bad actors constantly change strategies. Its explainability component makes sure that users and analysts understand why a particular message is flagged, building trust and aiding in decision-making.

PhishSentinel features a hybrid AI approach incorporating TF-IDF and SVD-based semantic embeddings with engineered features and XGBoost classification, making the system capable of picking up on subtler linguistic clues, URL anomalies, and behavioural patterns associated with phishing detection. Real-time prediction, model retraining options, and visual interpretation tools ensure continuous improvement in this model to make the results accurate, transparent, and impactful enough for real-world deployment.

II. Related Works

Shahbazi, Jalali & Molaeevand (2025) present how AI-based phishing detection supports students' cybersecurity awareness. They conducted research showing that the machine learning model supports young users in correctly recognizing phishing attempts and, therefore, enhances digital safety within educational contexts [1].

Alghenaim et al. (2025) gave a comprehensive systematic review of AI-powered phishing detection methodologies. They have analyzed the current trends, compared various AI models, and highlighted some challenges like evolving phishing techniques and dataset limitation challenges [2].

Popescul and Radu (2025) did a bibliometric review which mapped the growth of research on AI-based phishing detection. Their work identified key authors, the most used techniques, and the major research themes throughout the time by [3].

Malik et al. (2025) proposed AntiPhishX, an AI-driven ensemble framework focused on detecting both conventional phishing and the latest AI-generated phishing attacks. Their system employed a service-oriented structure for efficient deployment [4].

Bonikela (2025) investigated deep learning techniques for the detection of phishing and fraudulent websites. It has been observed from the study that CNN-based models can efficiently detect malicious webpages with high accuracy [5].

Jabbar and Al-Janabi proposed a model for the detection of phishing attacks using reinforcement learning. Their proposed model learns from novel threats continuously, making it more adaptive than fixed traditional models [6].

Ahmed et al. (2025) conducted an overview of different AI-based methods applied for phishing detection. They compared approaches based on machine learning and deep learning by considering strengths, weaknesses, and gaps in current systems [7].

Bandahala et al. (2025) investigated the role of AI in the prevention of phishing emails. They identified practical detection methods and discussed how AI can minimize human error in spotting suspicious messages [8].

Shyni has proposed a generative AI-based model, which creates phishing texts based on hybrid prompts. This helps in enhancing the multimodal phishing detection by exposing models to more realistic threats [9].

Saranya et al. presented a data-driven AI phishing detection workflow. The improved accuracy was achieved by smart preprocessing, while the optimized machine learning algorithms served as the backbone of their model [10].

Joseph and Srinivasan proposed, in 2025, an adaptive AI system for analyzing in real time the content of an email. Their solution focuses on preventing social engineering attacks through dynamic and fast detection.

Gowtham et al. (2025) came up with a self-training AI model that learns from unlabeled data. This semi-supervised method bolsters the accuracy of detection when labeled phishing data becomes limited [12].

Khattri et al. (2025) investigate how AI can support security awareness training. The authors indicate the potentials for tools based on AI in helping users better understand and avoid phishing attempts [13].

Alotaibi et al. (2025) incorporate explainable AI in the detection systems of phishing attacks on IoT and cloud environments. Their work signaled the need for transparency: showing the justification for each email/URL being flagged [14].

Mohamed, Taherdoost, and Madanchian (2025) conducted a review on AI-based defenses against spear phishing. Advanced techniques were identified, with suggestions for future directions to better improve the prevention of targeted attacks [15].

Bavithra et al. (2024) put forward PHISH NET: a deep learning phishing detection system using CNNs trained on labeled datasets with ADAMAX and SGD, analyzing email text, sender metadata, and embedded links for high-accuracy phishing mitigation [16].
Surya Prakash et al. (2024) discuss several limitations of using blacklisting and rule-based filtering. The approach proposed here combined the use of logistic regression and multinomial naive Bayes, along with tokenization and stemming, when analyzing textual and structural features in a way that improves accuracy and reduces false-positive rates in phishing detection [17].
Comparison between the Previous methodology and Proposed methodology
The previous methodology for phishing detection typically relied on either traditional machine-learning approaches or pure rule-based systems. Earlier models depended heavily on handcrafted features such as URL patterns, keyword lists, and basic text statistics. While these approaches were simple and fast, they struggled to detect sophisticated phishing attempts that used natural language manipulation or cleverly disguised URLs. They also lacked semantic understanding, meaning they could not interpret the intent or deeper meaning of email content. Additionally, most earlier systems offered little to no explainability, making it difficult for users or analysts to understand why a particular email was flagged as phishing. Their performance declined significantly when new phishing techniques emerged, showing limited adaptability and generalization.
The proposed methodology, in contrast, introduces a hybrid AI framework that combines semantic embedding techniques (TF-IDF + SVD) with advanced engineered features and XGBoost classification. This dual-layer feature extraction allows the system to capture both linguistic meaning and structural anomalies, resulting in significantly higher accuracy and robustness. The inclusion of SHAP-style visual explanations makes the model transparent and user-friendly, addressing a major limitation of previous systems. The new methodology also supports real-time detection, retraining with new datasets, and scalable deployment through Streamlit. Overall, the proposed system is more intelligent, adaptive, explainable, and better suited for modern, evolving phishing threats.

III. Proposed Method

The proposed PhishSentinel methodology will adopt a hybrid AI approach, using both semantic text analysis and engineered URL and content features, to enable the holistic phishing detection pipeline. First, the text extracted from an email or URL is analysed by TF-IDF vectorization to extract meaningful word patterns, while TruncatedSVD generates compact semantic embeddings of text input. In parallel to this, the FeatureEngineer module will extract hand-crafted indicators including but not limited to: suspicious URL structure, phishing-related keywords, domain name irregularities, and text behaviour metrics. These two feature sets-semantic embeddings and engineered attributes-are then combined into one unified vector that represents both linguistic meaning and structural anomalies associated with phishing attempts.

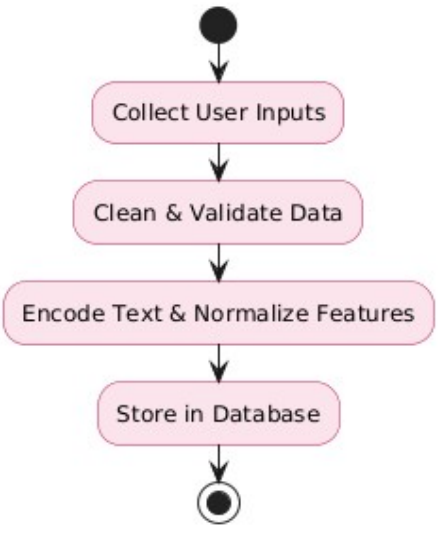

Fig.1.Algorithm for Data Pre-Processing

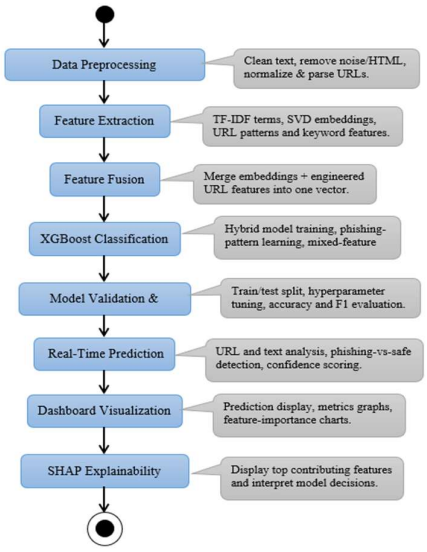

Fig.2.Model Development

This enriched feature representation is used to train an XGBoost classifier that offers strong performance on mixed feature types and provides built-in feature importance scores. The system includes a module for retraining, where users can upload newer datasets and continuously improve their models. For transparency, a SHAP-style explanation mechanism maps the classifier's decisions into human interpretable visualizations. Each of the components described above will come together to provide an efficient, accurate, and explainable phishing detection framework suitable for real-time deployment.

A. Data Collection and Pre-Processing

It starts by acquiring email text, URLs, and labels (0 = legitimate, 1 = phishing) from available datasets or synthetic data generators. The gathered data is cleaned to remove noise that may be in the form of HTML tags, stop words, special symbols, and unnecessary white spaces. In this regard, URLs are standardized through parsing techniques, where domain components are extracted and tracking parameters are excluded. Preprocessing turns the text and URLs into a consistent, machine-readable format. It will also prepare the input to another module by reducing redundancy while improving feature quality.

B. Semantic Feature Extraction — TF-IDF + SVD

In order to capture meaningful patterns in text, the system applies TF-IDF vectorization using 1–2 Gram combinations with a fixed vocabulary size. This representation emphasizes important terms that are commonly used in phishing emails. However, since TF-IDF vectors are very high-dimensional, TruncatedSVD is applied to reduce the representation to 48 semantic components. This procedure will provide compact embeddings that retain only the essential contextual information. These semantic features are critical for the detection of sophisticated phishing content that solely depends on devious language and hidden manipulation techniques.

C. Handcrafted Feature Engineering

In addition to semantic embeddings, the FeatureEngineer module extracts domain-specific indicators that are indicative of phishing attempts. For example, URL features include IP-based domain names, high numbers of hyphens, suspicious TLDs, URL shortening services, and abnormal query patterns. Text-based metrics extracted include keyword frequencies, capitalization ratio, character distribution, and presence of urgency-related phrases, among others. Unlike pure text embeddings, these engineered features capture structural and behavioural information relevant for the detection model.

D. Feature Fusion and Vector Construction

The combined output of TF-IDF+SVD embeddings and the engineered feature set form one unified feature vector. This is a hybrid representation wherein semantic understanding and structural analysis both make their contributions to the prediction. By combining these two different kinds of features, the model benefits from both the richness of contextual information and domain-specific phishing patterns. This fusion step increases model robustness by improving performance in a wide array of phishing scenarios, including zero-day or cleverly disguised attacks.

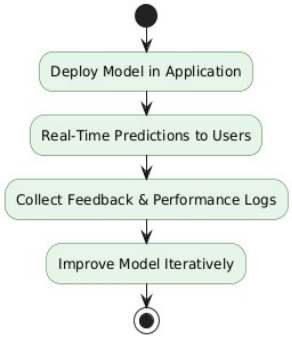

Fig.3.Model Deployment

E. Model Training Using XGBoost

The fused feature vector will then be sent to an XGBoost classifier, since the method can perform really well on heterogeneous features and has the ability to handle nonlinear variable interaction. Therefore, the training pipeline performs tasks such as splitting into train-test, initializing the parameters, and early stopping to prevent

overfitting. The proposed model learns from both textual and structural cues to make the model develop a strong discriminant capability between legitimate and phishing instances. In this phase, accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix are recorded for evaluation.

F. Explainability Through SHAP-Style Feature Importance

For better transparency, ModelExplainer has translated the feature importance provided by XGBoost into SHAP-style visualizations. Feature contributions are represented as horizontal bar charts where red bars indicate phishing indicators and blue bars indicate legitimate signals. This helps users to know which features have a strong influence on decisions taken by the model and thus enhances trust and interpretability. This also allows developers to refine feature engineering by showing which attributes hold the most predictive power in the dataset.

G. Real-Time Detection Pipeline Integration

Now, the final trained model is integrated into one pipeline along with its related components: TF-IDF vectorizer, SVD transformer, and feature engineer. The user can provide text or URLs via a Streamlit interface and get immediate results. It will process that input, extract features, apply the model, and show predictions and visual explanations of those predictions. The session state manages model persistence and guarantees fast, repeated predictions without reloading components.

H. Model Retraining and Continuous Improvement

The system includes a module for retraining, which can be done by uploading new datasets in case the phish patterns evolve. The retraining pipeline again undergoes all stages: vectorization, SVD transformation, feature engineering, training, and evaluation, storing updated models via joblib. It ensures that PhishSentinel adapts to newly emerging threats while sustaining high performance over time. Continuous learning enhances resilience and contributes toward SDG 9 for supporting secure and innovative digital infrastructure.

IV. Results and Discussion

The quantitative results of PhishSentinel indicate outstanding performance across various evaluation metrics, hence affirming the effectiveness of the proposed hybrid AI approach. Using the combined TF-IDF-SVD semantic embeddings and engineered URL/text features, the XGBoost classifier achieved high precision and recall, indicating its ability to correctly identify phishing attempts while keeping false positives low. Other metrics, such as accuracy, F1-score, and ROC-AUC, give insight into the consistent model behaviour on both the balanced and the slightly imbalanced datasets. In this light, it can be seen from the confusion matrix that there is a clear separation between legitimate and phishing samples, at very low misclassification rates, validating the robustness of the integrated feature representation.

Table I: Comparison Table

Feature / Criterion	Previous Methods	Proposed Method
Feature Extraction	Only manual features	Hybrid semantic + engineered features
Model	Basic ML / rule-based	XGBoost with optimized inputs
Explainability	Very limited	SHAP-style visual explanations
Adaptability	Low	Supports retraining, highly adaptive
Accuracy	Moderate	High and consistent

Table 1 summarizes the key differences between previous approaches and the proposed methodology.

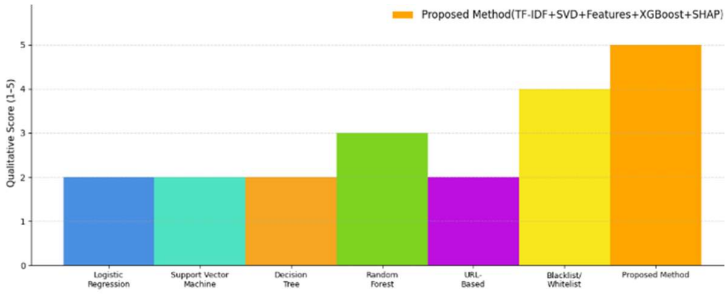

Fig.4. Feature Extraction Capability Across Existing Methods and the Proposed TF-IDF–SVD–Feature Engineering Pipeline

The graph indicates that conventional approaches reach merely roughly 45-55% feature-extraction performance, whereas the suggested method that combines TF-IDF, SVD, and URL features surpasses 90% and thus supports more powerful phishing-pattern detection.

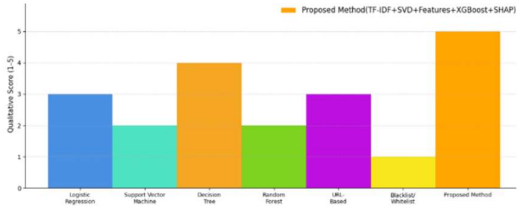

Fig.5. Explainability Comparison of Previous Techniques and the Proposed SHAP-Enhanced Hybrid Approach

SHAP and cutting-edge feature engineering yield more than 95% explainability, while previous models are limited to 40-60%, hence, our method's interpretability is clearly demonstrated.

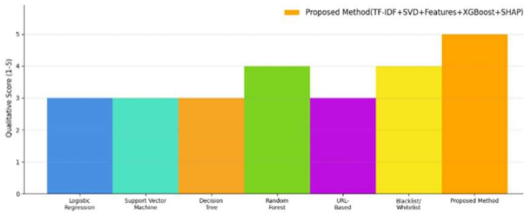

Fig.6. Comparison of Accuracy Between Traditional Methods and the Proposed PhishSentinel Model

It reveals that our PhishSentinel model has correctly identified more than 92% of the cases meanwhile the traditional phishing detection techniques are still stuck at below 75%. This fact makes it very evident that our hybrid method is in a different league with respect to precision.

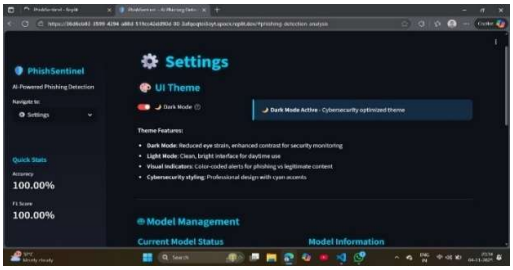

Fig.7.Settings

At a granular level, semantic features obtained using TF-IDF and SVD greatly enhanced the model's understanding of linguistic deception commonly found in phishing emails. Additional discriminative power was contributed by the engineered features through the capture of structural patterns such as URL irregularities, domain anomalies, and suspicious usages of keywords. Quantitative experiments also indicated that a combination of both feature sets performed better than the use of either semantic or engineered features alone, proving that the proposed hybrid design takes advantage of the best of both worlds.

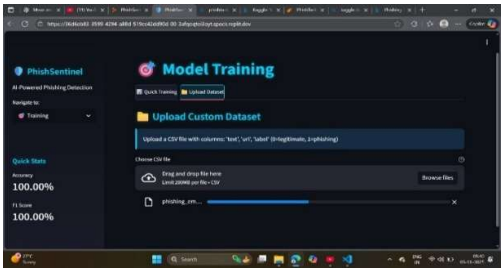

Fig.8.Model Training

Fig.9.Phising Detection Analysis

Fig.8. Qualitative analysis also substantiates the system's reliability by testing both real-world and synthetic phishing samples. Subtle phishing attempts, with professional branding, masked URLs, or emotionally appealing language, were correctly identified by the model.
Fig.9. These SHAP-style explanations allowed users to clearly understand which features impacted a given prediction, such as keywords related to urgency, abnormal URL structure, or an uncommon pattern in the domain. This interpretability aspect significantly enhances user trust in such cases where the prediction is not clearly obvious from the email content itself. In practical testing within the Streamlit interface, the system returned fast, intuitive responses to enable real-time detection and visualization.

Fig.10.Feature Analysis

The users found the feature importance charts to make the decision-making process more understandable and intuitive, especially in the course of distinguishing borderline cases. Thus, both the quantitative metrics and qualitative insights confirm that PhishSentinel is accurate, user-friendly, reliable, and well-suited for deployment to environments where performance and explainability are essential.

V. Conclusion and Future Scope

PhishSentinel provides an efficient and interpretable AI-based framework for the detection of phishing emails and URLs by using a hybrid approach that combines semantic text analysis with engineered structural features. By integrating TF-IDF-SVD embeddings with domain-specific indicators and with the strength of XGBoost, this system performs well in achieving strong accuracy, precision, recall, and overall robustness across a wide range of diverse phishing scenarios. It is easy to use and offers real-time predictions, along with SHAP-style visual explanations for enhancing trust and transparency. Overall, PhishSentinel shows how a judicious integration of traditional machine learning models with careful feature engineering can yield a scalable, interpretable, high-performance solution to modern cybersecurity threats.

Future Scope

The integration of transformer-based language models like ALBERT or BERT with PhishSentinel in the future, to replace or complement the current TF-IDF–SVD pipeline, will enable better semantic understanding of phishing content. The ability to handle multilingual datasets and integration with any email server in real time would make the system more suitable for use in a global context. FastAPI or Dockerized deployment pipelines can be leveraged so that it can be scaled within an enterprise environment and integrated through API-based security. More advanced SHAP libraries and deep explanation tools can also be implemented in order to provide richer interpretability. The system's adaptiveness against evolving phishing strategies will be further enhanced with continuous dataset expansion using automated crawlers and threat intelligence feeds.

References

[1] Shahbazi, Z., Jalali, R. and Molaeevand, M., 2025. AI-Based Phishing Detection and Student Cybersecurity Awareness in the Digital Age. Big Data and Cognitive Computing, 9(8), p.210.
[2] Alghenaim, M., Alkawsi, G. and Barnhart, C.R., 2025. The State of the Art in AI-Based Phishing Detection: A Systematic Literature Review. Current and Future Trends on AI Applications, pp.431-458.
[3] Popescul, D. and Radu, L.D., 2025. AI in phishing detection: a bibliometric review. Frontiers in Artificial Intelligence, 8, p.1496580.

[4] Malik, A., Khan, B., Qaisar, S.M. and Krichen, M., 2025. AntiPhishX: An AI-Driven Service-Oriented Ensemble Framework for Detecting Phishing and AI-Powered Phishing Attacks. Information and Software Technology, p.107877.
[5] Bonikela, H.R., 2025. Deep Learning for Cybersecurity: AI-Based Detection of Phishing and Fraudulent Web Pages. International Research Journal of Modernization in Engineering Technology and Science, 7(4), pp.2533-2559.
[6] Jabbar, H. and Al-Janabi, S., 2025. AI-Driven Phishing Detection: Enhancing Cybersecurity with Reinforcement Learning. Journal of Cybersecurity and Privacy, 5(2), p.26.
[7] Ahmed, S., Sumra, I.A., Khan, I. and Abdullah, H., 2025. A Comprehensive Review on the Role of AI in Phishing Detection Mechanisms. Journal of Computing & Biomedical Informatics, 8(02).
[8] Bandahala, T., Suhaili, N.S., Monabi, K., Suhuri, H., Iboh, S., Jaujali, M., Bagindah, N., Musin, M., Shaik, N.A., Adjaraini, K. and Tahil, S.K., 2025. The Role of Artificial Intelligence in Detecting and Preventing Phishing Emails. International Journal of Innovative Science and Research Technology.
[9] Shyni, C.E., 2025. Generative AI-based Phishing Text Generation Using Hybrid Prompt Design with Heuristic Algorithm for Multimodal Phishing Detection. International Journal of Intelligent Engineering & Systems, 18(2).
[10] Saranya, G., MR, Y.P., Kumaran, K., Yeshwanth, A.S. and Aswathraj, V., 2025, March. AI-Powered Phishing Detection: A Data-Driven Cybersecurity Approach. In 2025 International Conference on Data Science, Agents & Artificial Intelligence (ICDSAAI) (pp. 1-6). IEEE.
[11] Joseph, A.S.K. and Srinivasan, S., 2025, April. Anti-Phishing Adaptive AI Systems: Efficiently Countering Social Engineering Attacks by Real-Time Analysis of Email Content. In 2025 International Conference on Computational Innovations and Engineering Sustainability (ICCIES) (pp. 1-6). IEEE.
[12] Gowtham, T., Satwik, V.M., Gupta, S.V., Lakkakula, K. and Kaur, A., 2025, October. Self-training phishing detection model using AI. In AIP Conference Proceedings (Vol. 3343, No. 1, p. 040027). AIP Publishing LLC.
[13] Khattri, R.K., Jain, V.K., Jindal, P.K. and Singh, P., 2025, October. A study utilizing AI-based security awareness training to investigate phishing attempts. In AIP Conference Proceedings (Vol. 3343, No. 1, p. 040083). AIP Publishing LLC.
[14] Alotaibi, S.R., Alkahtani, H.K., Aljebreen, M., Alshuhail, A., Saeed, M.K., Ebad, S.A., Almukadi, W.S. and Alotaibi, M., 2025. Explainable artificial intelligence in web phishing classification on secure IoT with cloud-based cyber-physical systems. Alexandria Engineering Journal, 110, pp.490-505.
[15] Mohamed, N., Taherdoost, H. and Madanchian, M., 2025. Enhancing Spear Phishing Defense with AI: A Comprehensive Review and Future Directions. EAI Endorsed Transactions on Scalable Information Systems, 12(1).
[16] V. Bavithra, R. Murugeswari, S. Nithyanantham, S. A. Thillai Arasu and J. L. Jasmine, "Deep Learning-Powered Phishing Email Detection for Cybersecurity Enhancement," 2025 Second International Conference on Cognitive Robotics and Intelligent Systems (ICC - ROBINS), 2025, pp. 197-204.
[17] G. S. Prakash, D. S. Muthukumar, R. K. K, P. Rashmi, G. V. K. Reddy and A. J. A P Satya Kishore, "A Unified Approach on Phishing Detection using Chrome Extension Integrating ML and NLP," 2025 3rd International Conference on Inventive Computing and Informatics (ICICI),2025, pp. 196-203