Grenze International Journal of Engineering and Technology, June Issue

SafeNet: AI-Powered Browser Extension for Real-Time Phishing Site Detection, user Protection, and Instant Warning Alerts

Jenila Nilofar J[1] and Denisha M[2]
[1-2]Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore, Tamil Nadu,
Email: jenilanilofar@karunya.edu.in, denisha@karunya.edu

***Abstract*— Of particular concern in terms of cyber attack is phishing and other hackers because they are busy adapting URLs, domain configuration and deceptive content patterns, as a means of making it off-the-shelf blacklists and rule sets. The paper presents a phishing detection system (SafeNet) built on AI that works by examining URLs on-the-fly and can issue a warning to the computer user. It involves a combination of URL and domain-oriented features record with the aided machine learning models, including Neural Network, K-Nearest Neighbors, Naive Bayes, random forest and Support Vector Machine. The preprocessing, as well as the choice of the model is also performed with the help of Principal Component Analysis and Recursive Feature Elimination to enhance the quality of the features and remove redundancy. Its back-end prediction is based on FastAPI platform and it has a real time front-end interface based on URL submission and classifying. With experimental comparison of multiple classifiers, the model of the Neural Network displays the highest overall performances as compared to the methods being tested and is more accurate, more precise, recalls more, and F1-score better. According to the study, the main value of DanceNet is not only that it integrates multi-model phishing detection, feature optimization, and deployable real-time prediction in a single workflow, but also offers this workflow as a service. The findings show that the suggested framework can be used to facilitate the feasible phishing detection as well as to offer an opportunity to expand it in the future via ongoing feedback based enhancement.**

***Keywords*— Concepts Phishing Detection, Machine Learning, Neural Network, K-Nearest Neighbor, Naive Bayes, Random Forest, Support Vector Machine, Principal Component Analysis, Recursive Feature Elimination, Feature Selection, Web Scraping, WHOIS, FastAPI, Streamlit, Cybersecurity, URL Classification.**

I. INTRODUCTION

One of the highest proliferations of cyber-attacks today belongs to Phishing Attacks where a person in a position of power uses a lot of finances, usernames, passwords and finances by impersonating to gain access to important information to obtain this data. Hacking is a major problem yet the technology dedicated to its security is left in the hands of people through frauds, malicious websites and social engineering solutions. The nature of phishing attack is never similar and this is what renders the rule based type of detection systems impotent to keep up with the changes that have occurred.

Grenze ID: 01.GIJET.12.2.623_26

SafeNet is developed based on a realistic detection process where the extracted attributes are URL-based attributes and domain-based attributes that are processed to classify phishing. The framework takes into account aspects like the length of URL, suspicious lexical habits, domain age, and information related to registration, as well as traffic-based indicators. These characteristics are then provided to several monitored learning models to enable a comparison of the behavior of classifiers instead of expecting that there is one single model that will address all phishing patterns.

Competent use of machine learning to detect phishing is not the primary contribution of the work, but uniting three components into a single deployable environment: multi-model comparative classification, use of PCA and RFE to filter out features and predict in real-time through an API-based structure. Moreover, the new system is designed to be interactive and applicable in the future with the adoption of feedback. This blend enhances the practice applicability of the work and contrasts it with studies, which only conceptualize the detection of phishing or only experiment phishing recognition, but not practically.

This work has the following contributions. First, SafeNet offers a single phishing detection process that integrates the extraction of features, pre-processing, comparison of classifiers and real-time prediction aid. Second, the analysis conducts comparative analyses between Neural Network, KNN, Naive Bayes, Random Forest, and SVM models rather than using one classifier. Thirdly, the impact of feature selection and dimensionality reduction on detecting performance is analyzed with the help of PCA and RFE. Fourth, the structure includes deployment using a Fastapi backend and user-facing interface, thus enhancing more practical usability as compared to offline application.

II. Literature Survey

Even more recent research only made it even more evident that the use of artificial intelligence (AI) and machine learning does become appropriate when it comes to phishing attacks detection and prevention and the question of cybersecurity, overall. It is possible to use the machine learning models and help AI to identify a taboo Web site and a rogue email as Lamina et al. [1] depicted. In their article, they ruminate the merits of properly developed AI algorithms to discover the evidence of phishing and, therefore, streamline the preventive and reactive processes. Equally, Prince et al. [2] highlighted the initiatives of the AI-based methods that offer the utilization of the data to boost the quality of threat detection and its speed of counteractions. Through their research, it was evident that AI systems could process large amount of data in the actual place and react swiftly to the cyber-attacks by detecting them. Following what was described by Lamina et al. [1]; one can utilize machine learning models and AI can assist in discovering a phishing site and a counterfeit email concerning the trickery. They conclude that AI will be able to detect the anomaly in the pattern of activity that might occur and therefore, it will be possible to detect the potential cyberattacks early enough and prevent it. The other concept that Ali [4] has explored was securing a massively scaled system against phishing and intrusion attacks with the assistance of supervised machine learning classifiers. The mass data processing has concluded that the AI models can be used to determine the legitimate actions and the malicious actions so as to enhance the safety of the systems.

Negotiations concerning the development of phishing attack with the help of AI based browsers were underway between Arun and Abosata [5]. Their work has also recommended that as the attackers improve how they develop better phishing attacks with the help of AI, the more etiquettes will be employed in identifying and curbing them. Another broader application of AI in cyber-security that Naseer [6] also explained could be deployed to determine a possible security breach and trend of a cyber-attack.

This has extensively been applied on other AI-based applications in utilizing phishing identification with fraud prevention and digital identity protection. Banu [7] touched on AI based technologies of safeguarding online identities, particularly the online transaction do per se. The aspect of identity verification using AI has been highly important at least as far as the minimization of fraud is concerned given the trend of digital payments is under a continually accelerating spree. [8] discussed the issue and trends in the systems of AI-based phishing detection, and provided the weaknesses of the currently used systems, as well as additional opportunities in the further progress of phishing detection systems that will lead to a higher accuracy and reliability of the systems. It has been further pointed out that it needs to embrace AI in the new techno systems such as Internet of Things (IoT). The issue of AI-based phishing detection introduced in smart home by Fatima et al. [9] has been eliminated and the need to mitigate cybercrimes on IoT-based devices. Malik [10] examined the existing vulnerability forecasting and analysis of the attack vectors that relies on AI that can be leveraged to determine how preventative analytics is relevant to organizations and safeguard against security risks before they occur.

Kumar [11] too studied the next-generation phishing detection system by the AI, and contemplated on the further advanced systems that could easily identify the genuine and the malicious content.
More studies have been conducted in the use of AI in larger cybersecurity devices. Muthusamy [12] studied the fact machine learning was the motor that instigated the threat detection systems which provided analyses to large amount of security informational data in their real time. Ansarullah et al. [13] studied the AI-based antimalware tools and have shown how deep learning and neural networks have the potential to be used to complement the existing malware detection tools. The feasibility of AI is also regarded in the study within the financial aspect of cybersecurity since the articles by Shabir and Khalid demonstrate that it can be applied to the sphere of fraud detection successfully to interpret the patterns of the transactions. Finally, Kota [15] has suggested an artificial intelligent system of fraud detection where it is micro-architecture hence enhancing the availability of scalability and responsiveness of financial security systems. Generally, as depicted in the literature, AI-based phishing detection has evolved more of a rule-based filtering to a data-based classification and real-time threat analysis. Nonetheless, in the existing literature, much of concern is given on the level of model precision or the oratory distinction of AI in cybersecurity, whereas the comparatively minor topics on feature engineering, comparative classifier assessments, and deployable real-time system structure exist within a single framework. SafeNet can be placed in this context as implementation-based research which unites these points and measures them together in a phishing URL detection piping.

III. Proposed Methodology

Fig. 1 shows the workflow of the SafeNet system in input URL to the final phishing. The methodology is divided into four significant steps, namely, feature extraction, preprocessing, model training and comparison, and real-time prediction support. This design is based on the need to make the system not restricted to checking of the URLs in the system but rather do systematic feature-based classification of the system with various supervised learning models. It has PCA and RFE to help minimize redundant features in the data set and keep the most informative attributes to train the model. This was the methodology chosen to enhance predictability as well as explainability of the process of the classification.

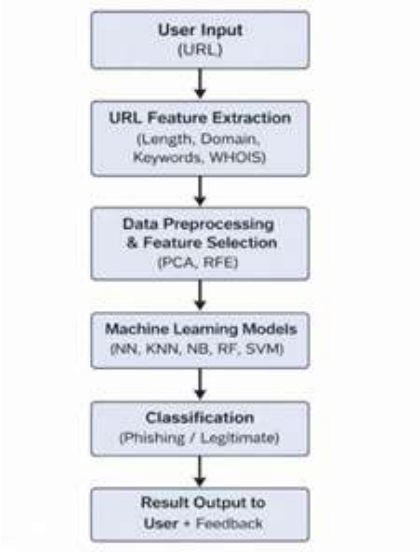

Figure 1: Flowchart

The Using web scraping and whodis, the system is sensitive to capture dynamic URL based features that encompass length of the URL, suspicious keywords, age of the domain and registration period and web traffic. It also uses FastAPI as the back-end which will carry out the live predictions and has a Streamlit experience which will interface with the users and provide the feedback.

A. Data Preprocessing and Collecting.

Phishing workflow is based on the process of URL records (phishing or legitimate) which are classified and labeled before use by the workflow. Once data is collected, it is preprocessed to make the extracted attributes

ready to go through machine learning. This phase involves the processing of untidy values and the encoding of categorical data where it occurs and deserializing numeric data with StandardScaler. Because the features space is numerically similar between the feature combinations, and less sensitive to scale variation, it requires preprocessing since the methodology of feature modeling is essential toperformance with a limited range of feature combinations.

Two refinement strategies of features are introduced to enhance the efficiency of models further. Dimensionality reduction is based on the Principal Component Analysis because the correlated information of the features can be effectively represented in a smaller space. The Recursive Feature Elimination approach is employed on finding the most informative subset of features in order to classify. The combinations of these steps are aimed at reducing noise, eliminating less significant properties, and facilitating better generalization of various patterns of phishing.

Accuracy, precision, recall and F1-score based on confusion matrix are used to evaluate. The general correctness of prediction is given by accuracy, the probability of making predictions correctly is given by precision, the capacity to find phishing is given by the recall, and the F1-score is a compromise given by the precision and recall. The reason why these metrics are employed is that phishing detection is not just an issue of accuracy, but also one of security whereby, any miss detection as well as false alarms should be minimized.

Evaluation Metrics are To determine the performance of phishing detection models that give a thorough account of phishing performance they are also ascertainment objectives like accuracy, Recall and F1-score based on confusion matrix.

$$\text{Accuracy} = \frac{TP+T}{TP+TN+FP+FN}.$$

Precision is defined as the number of correctly discovered phishing URLs at a quotient of the number of URLs selected as a phishing URL: $\text{Precision} = \frac{TP}{TP+F}$,

Recall (also known as Sensitivity) In the model's ability to detect phishing URLs correctly: $\text{Recall} = \frac{TP}{TP+FN}$,

F1-Score is the harmonic mean of the Precision 2. Finally, Recall is providing a balanced performance measure.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Where:
TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

B. Feature Engineering

SafeNet feature engineering looks at deriving quantifiable URL and domain features that can be used to differentiate phishing pages and legitimate web pages. They involve lexical features like URL length, structure of hostname, special characters, and indicative key words and domain features like domain age, registration features and traffic related information. In the choice of these features, the motivation comes in the fact that, frequently, the phishing websites have dissimilar counterparts on the structure, naming conduct and domain history. On extraction, the features are analyzed by PCA and RFE in such a way that the resulting feature space on which training is carried out will be relevant as opposed to quantifiable. This enhances the strength of the learning process and less reliance on redundant qualities. Consequently, feature engineering in SafeNet is not considered as a distinct preprocessing stage in isolation rather than a key part of the detection approachology.

C. AI Model Development

SafeNet uses a strategy of comparative model development whereby multiple trained supervised classifiers are trained and evaluated on the same feature pipeline. The models chosen are Neural Network, K-Nearest Neighbors, Naive Bayes, Random Forest and Support Vector Machine. This option enables the research to make comparisons among nonlinear learning behavior, distance-based classification, probabilistic prediction, ensemble learning and margin-based classification of the same phishing detection problem.

Learners near and Among these are the Neural Network model as the main deep learning classifier as it can learn intricate associations among extracted phishing indicators. The rest of the classical machine learning models are provided to give a considerable background and a point of reference. The evaluation metrics used in model performance are the same ones used to evaluate performance of the model so that the decision on the model used

is arrived at on the basis of evidence that can be measured as opposed to an assumed one. This design enhances the rigour of the methodology of the study and helps in identifying the most promising classifier to be deployed.

D. live Detection and API Integration.

SafeNet is designed as a phishing detections architecture to go beyond offline experiments. The trained model is run with a Fast API backend which is reached by a user-supplied URL, makes the needed feature processing, and emits either a phishing or a legitimate prediction. An interface that is easy to use is provided to allow interactive analysis and simplify the framework to be used in realistic environments.

This integration is critical since the phishing detection systems would be most effective when it is able to perform in real-time in contrast to being research prototypes. This fact-driven deployment design of SafeNet is thus a significant component of the methodology, tied with the model development and realistic cybersecurity application.

Dataset Description involves the previously equal phishing and legit URLs that were collected in the PhishTank, OpenPhish and the trusted sources of web. Extracted features include URL scraping, scraping of web pages and the WHOIS searches. The preprocessing operations produced prior to the training and analysis of the models involve expertizing, codification and option of features.

TABLE I: DATASET CHARACTERISTICS

Attribute	Description
Data Source	Phish Tank, OpenPhish, and trusted legitimate websites
Total URLs	Phishing and legitimate URLs combined
Class Labels	Phishing, Legitimate
Feature Type	URL-based, Domain-based, Content-based
Number of Features	Multiple extracted features (before PCA/RFE)
Feature Extraction Methods	URL parsing, Web scraping, WHOIS queries
Preprocessing Techniques	Missing value handling, normalization, feature selection
Dimensionality Reduction	Principal Component Analysis (PCA)
Feature Selection	Recursive Feature Elimination (RFE)
Learning Type	Supervised Machine Learning

E. Methodological Justification

The approach taken in SafeNet is driven by the fact that three of the practical requirements of phishing detection, which include feature diversity, classifier reliability, and deployment feasibility are aimed to be addressed. One model, that has been trained on raw features, might work well on one dataset but not generalize in case of phishing patterns changes. This is why SafeNet uses isolated feature extraction made in conjunction with PCA and RFE and assesses a variety of classifiers in a common pipeline. The design enables the study to not only determine the best performing design, but also determine the impact the refinement of features has on the quality of prediction. Moreover, there is a real-time deployment layer since phishing detection cannot be practically used when it is in-offline mode only. The interface and Fast API interaction consequently assist in translation of the model output to useful warning support. This integration of feature processing and comparative learning and deployment support is the methodology of the proposed system.

IV. RESULTS AND DISCUSSION

The experimental analysis compares the performance of the phishing detection of Neural Network, K-Nearest Neighbors, Naive Bayes, Random Forest and Support Vector machines models with a shared feature engineering and preprocessing pipeline. It is compared in terms of URL and domain-related characteristics such as suspicious pattern of lexics, the information of domain registration, domain age, and indicators reflecting traffic. The refinement of the feature space before classification is done using PCA and RFE. The aim of the results section is thus not merely to provide performance values, but to present a demonstration of the capability of providing comparative modelling as well as feature refinement to the general SafeNet frame.

A. Model Analysis of performance.

The experimental findings indicate the evident divergence in the effectiveness of the classifiers when it comes to the task of detecting phishing. The models that were tested were able to achieve optimal performance with a score of 97.4 on accuracy, 96.8 on precision, 97.2 on recall, and 97.0 on F1-score. These values suggest that the model can help capture the interactions between features of phishing better than the other approaches tested.

Random Forest and SVM also performed well, and their performance has been consistent, which implies that the ensemble and margin-based classification can be applied in the extracted feature set. The two were however less effective than the Neural Network. KNN and Naive bayes demonstrated reduced recall and F1-score implying lesser flexibility to phishing patterns as depicted in the data. The comparison allows concluding that the chosen feature pipeline can be employed with multiple models, whereas the Neural Network is the most appropriate final classifier of the models tried in the current paper.

TABLE II: MODEL PERFORMANCE COMPARISON

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
NN	97.4	96.8	97.2	97.0
KNN	81.0	81.5	77.0	79.1
Naive Bayes	76.0	75.2	70.5	72.8
Random Forest	93.0	91.0	92.5	91.7
SVM	92.0	89.7	91.0	90.3

B. Feature and Error Analysis

The visual analysis supports the numerical results by demonstrating the role of some characteristics in discrimination in phishing. This boxplot of the URL length v/s target class shows that phishing URLs are more widely dispersed and in most cases have a longer structure, compared to legitimate URLs. This validates that the URL length is a relevant discriminative characteristic in the SafeNet pipeline, however, not alone and should be straddled with additional lexical and domain based indices.

The confusion matrix of the Neural Network model also demonstrates the majority of the phishing and legitimate cases are being positively classified with the number of misclassifications being relatively low. This corroborates the quantitative finding that the Neural Network has a low false positive and false negative rate when compared to that of the other models. These visualizations are more powerful in the interpretation of the results, not serving as illustrations.

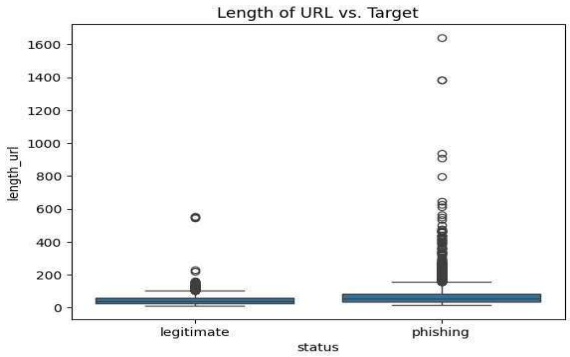

Figure 2: Length of URL vs. Target

Fig. 2 The boxplot clearly shows that the length of the URL is a distinguishing feature, with phishing URLs exhibiting wider variability in length compared to legitimate URLs.

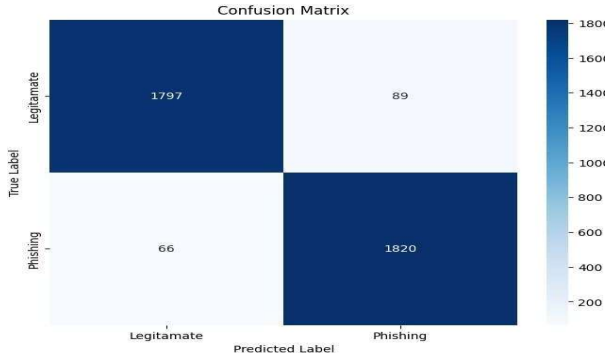

Figure 3: Confusion Matrix

Fig. 3 shows the outcome of the perplexation as or division of the True Positives, True Negatives,False Positives and the False Negatives of the phishing and the legitimate URL classification of Neural Network (NN) model. The result findings reveal that NN model is effective to obtain high true positive and low mis classification errors.

C. False Positive and Negative rates.

False Negative Rate (FNR), and False Positive Rate (FPR) were investigated to further estimate the reliability of each of the classification model.s. With the Neural Network (NN) model, the smallest FPR and FNR were achieved; however, it implies that the model is useful to control the incorrect classification of a phishing site, and even legitimate URLs.

On the other hand, FPR and FNR were the highest with Naive Bayes model, suggesting that it has the tendency to under-classify the URLs. The error rates of the models of the Random Forest and the SVM were relatively low; however, the models were still worse compared to the NN model.

TABLE III: FALSE POSITIVE AND FALSE NEGATIVE RATES

Model	False Positive Rate (%)	False Negative Rate (%)
NN	2.6	2.8
KNN	18.0	16.5
Naive Bayes	24.0	27.0
Random Forest	7.0	6.5
SVM	8.0	6.0

D. False Positive and False Negative Analysis

In False negatives are especially paramount in phishing detection since, in this case, they amount to malicious URLs false positive reduces usability as the safe URLs are incorrectly labeled. This is why the false negative rate and the false positive rate will give a valuable complement to the evaluation based on accuracy.

Neural Network model had lowest error rates compared to the other compared classifiers with a false positive and false negative rates of 2.6% and 2.8, respectively. This means a balanced performance of security orientation. Random Forest and SVM also had relatively low levels of errors, although KNN and Naive Bayes resulted in significantly higher ones. These results establish that the most competent models are not only more precise in general, but also more dependable to reduce prediction errors with detrimental impacts that are in practice.

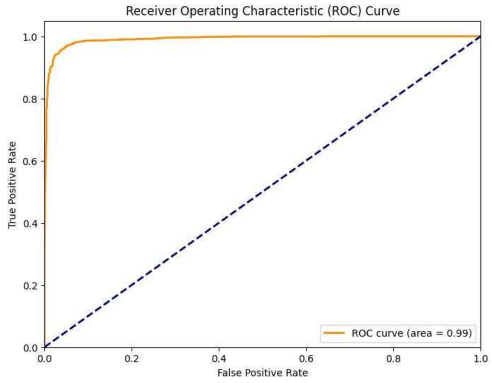

Figure 4: Receiver Operating Characteristic (ROC) Curve

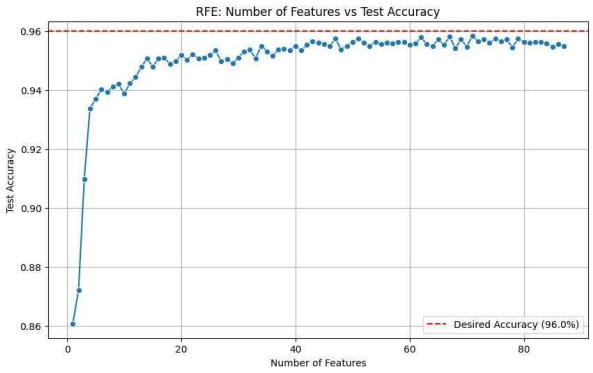

Figure 5: RFE: Number of Features vs. Test Accuracy

Fig. 4 below represents this ROC curve which represents the True Positive Rate as a function of False Positive Rate indicating the capacity of the model to accurately make a classification of the phishing URLs. The AUC (Area Under Curve) is 0.99 indicating that it is an excellent performance.

This plot Fig. 5 demonstrates how Recursive Feature Elimination (RFE) impacts model performance as the number of features selected increases. It shows a clear trend that as more features are selected, the accuracy approaches its maximum value of 96% .

This chart Fig. 6 compares the time required for PCA fitting and Random Forest training as the number of PCA components increases. It shows the computational cost of applying PCA and training the model, emphasizing the impact of dimensionality reduction on processing time.

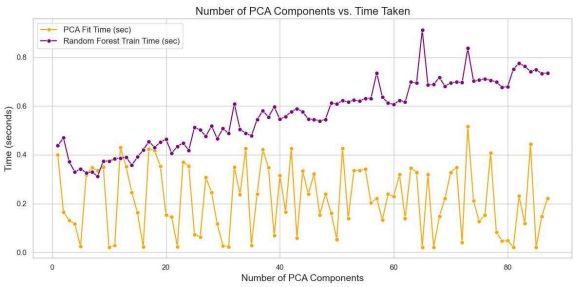

Figure 6: Number of PCA Components vs. Time Taken

E. Discussion of Findings and Limitations

The findings back the argument which informs about an efficient phishing detection framework in case of the Neural Network classifier which is accompanied with well-filtered feature selection and real-time deployment assistance. Meanwhile, the study is limited. This analysis of what as report is made it focus on the reported dataset and the chosen features and testing on a wider basis with more varied and changing phishing sources would also enhance generalizability. Besides that, future work has the capacity to give developed latency, scalability and user-feedback evaluation although the actual time architecture is documented in the deployment environment. By accepting such limitations, the study will be more reliable and will make the assertions more consistent with the evidence given.

V. CONCLUSION

The author presented a phishing detection model, SafeNet, in this paper, which is an AI-based model that incorporated feature extraction, preprocessing, comparative model evaluation, and support prediction based on real time. This study has experimented with different classifiers, i.e. Neural Network, KNN, Naive Bayes, and Random Forest as well as SVM in a standard phishing detection pipeline with the help of PCA and RFE. The experimental results proved that the best overall performance of the Neural Network was observed in terms of the accuracy, precision, recall, F1-score and reduction of mistakes.

It is mentioned in the article that feature refinement, model comparison can be used to improve phishing detection along with deployment-oriented system design. The contribution of the SafeNet, in this aspect, is not only in the accuracy of classification, but it offers a framework of practicality, which can be exploited in offering warning support on real time. The larger datasets, the larger number of comparative experiments, the adaptive learning with the assistance of the feedback, and much more deployed testing in the conditions of a real-life browsing can further develop the future work.

REFERENCES

[1] Lamina, O. A., Ayuba, W. A., Adebiyi, O. E., Michael,G. E., Samuel, O. O. D., & Samuel, K. O. (2024). Ai-Powered Phishing Detection And Prevention. Path of Science, 10(12), 4001-4010.

[2] Prince, N. U., Faheem, M. A., Khan, O. U., Hossain, K., Alkhayyat, A., Hamdache, A., & Elmouki, I. (2024). AI-powered data-driven cybersecurity techniques: Boosting threat identification and reaction. Nanotechnology Perceptions, 20, 332-353.

[3] Sadaram, G., Karaka, L. M., Maka, S. R., Sakuru, M., Boppana, S. B., & Katnapally, N. (2024). AI-Powered Cyber Threat Detection: Leveraging Machine Learning for Real-Time Anomaly Identification and Threat Mitigation. MSW Management Journal, 34(2), 788-803.

[4] Ali, H. (2022). AI-Powered Supervised Classifiers in Big Data Environments for Phishing Defense and Intrusion Detection.

[5] Arun, A., & Abosata, N. (2024). Next Generation of Phishing Attacks using AI powered Browsers. arXiv preprint arXiv:2406.12547.

[6] Naseer, I. (2024). The role of artificial intelligence in detecting and preventing cyber and phishing attacks. European Journal of Advances in Engineering and Technology, 11(9), 82-86.

[7] Banu, A. (2024). AI-Powered Digital Identity Protection: Preventing Fraud in Online Transactions.

[8] Kwaku, W. K. (2022). AI-Powered Phishing Detection Systems: Challenges and Innovations. Advances in Computer Sciences, 5(1).

[9] Fatima, N., Ashraf, M., Tehseen, R., Omer, U., Sabahat, N., Javaid, R., ... & Zaheer, A. (2024). AI-Powered Phishing Detection and Mitigation for IoT-Based Smart Home Security. Journal of Computing & Biomedical Informatics, 8(01).

[10] Malik, S. (2024). AI-Powered Cyber Risk Assessment: Predicting Vulnerabilities and Attack Vectors in Real-Time.

[11] Kumar, A. (2023). Next-Generation Approaches to Detecting and Preventing AI-Generated Phishing Scams. Eastern European Journal for Multidisciplinary Research, 2(1), 83-99.
[12] Muthusamy, K. (2025). AI-Powered Threat Detection in Cybersecurity Infrastructures. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 1(01), 24-33.
[13] Ansarullah, S. I., Wali, A. W., Rasheed, I., & Rayees, P. Z. (2024). AI-powered strategies for advanced malware detection and prevention. In The Art of Cyber Defense (pp. 3-24). CRC Press.
[14] Shabir, G., & Khalid, N. AI-Powered Fraud Detection and Risk Assessment: The Future of Financial Services.
[15] Kota, A. (2024). REAL-TIME AI-POWERED FRAUD DETECTION: A MICROSERVICES APPROACH. Technology (IJCET), 15(6), 2011-2024.