SafeNet: AI-Powered Real-Time Phishing Detection and Prevention System

Nithees M
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu, India
Email: cb.en.u4cse21338@cb.students.amrita.edu

Preethi V
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu, India
Email: cb.en.u4cse21365@cb.students.amrita.edu

Polina Likhitha
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu, India
Email: cb.en.u4cse21246@cb.students.amrita.edu

Yashwanth Ram
Amrita Vishwa Vidyapeetham
Coimbatore, Tamil Nadu, India
Email: cb.en.u4cse21370@cb.students.amrita.edu

Abstract – **Cybersecurity defense against phishing attacks is an important issue and the techniques of the traditional ones often are not enough to stay one step ahead of the evolving tactics. SafeNet is an AI based phishing detection system that uses Neural Networks (NN), K Nearest Neighbors (KNN), Naive Bayes, Random Forest, Support Vector Machine (SVM) and Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) to enhance feature selection and dimensionality reduction. Web scraping and WHOIS queries give the system an ability to extract comprehensive features from URLs such as URL length, domain age, traffic data and suspicious keywords. Performance metrics like accuracy, precision, recall, and F1 score are used in the evaluation of the models and we found the NN model to outperform the other models. Real time URL classification is achieved through deployment of a system on a FastAPI backend and a Streamlit front end, allowing users to submit URLs and get immediate feedback. In order to provide robust and scalable protection against phishing attacks, we design the approach to constantly adapt to the new threats with continuous user feedback.**

Keywords— Phishing Detection, Machine Learning, Neural Networks, K-Nearest Neighbors, 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

Phishing attacks are currently one of the most prevalent cyber threats that target users by posing as trustworthy entities in order to obtain mediums such as usernames, passwords along with financial data. Phishing continues to be a significant security challenge, even with security technologies having advanced to the point that it has accommodated for the continuous evolution of these attacks. Attackers exploit human error and trick users into giving up confidential data by using all manner of deceptive emails, fake websites, and social engineering tactics. Phishing techniques are becoming increasingly sophisticated and classical rule based methods have fallen behind; hence increasingly sophisticated techniques are required.

Sensitive cybersecurity issues such as phishing detection have been addressed very well through Artificial Intelligence (AI) and Machine Learning (ML). These technologies permit systems to learn from the history data and identify patterns that represent attempts to phishing, making accuracy and adaptability better. However, the heterogeneous nature of approaches employed in phishing poses a problem in that phishing detection needs to be a multi faced problem, since there are multiple machine learning models that need to be combined to ensure precision. In this paper, we present SafeNet, a complete AI based phishing detection system that uses a set of machine learning algorithms and methodologies to detect and stop the phishing attacks on time.

To address the problem, SafeNet uses several state of the art machine learning models including Neural Networks (NN), K-Nearest Neighbors (KNN), Naive Bayes, Random Forest and Support Vector Machine (SVM) to evaluate various parts of a URL and predict whether it is phishing or not. Using Principal Component Analysis (PCA) and Recursive Feature

Elimination (RFE) to reduce the dimensionality as well as select features, the performance of these models is further increased. These techniques cut out the low value features while leaving in the most available features thus lowering down the time and accuracy of models mainly when dealing with large data sets.

SafeNet feature extraction process involves a number of URLs attributes features generated by URL including length of URL, presence of suspicious keywords (e.g., login, bank, update), age of the domain, length of registration, as well as web traffic data. However, their presence here is very critical since phishing websites show distinct characteristics, which are not the case with legitimate websites. Furthermore, SafeNet can dynamically fetch such information as domain registration data as well as Alexa traffic rank through WHOIS and scraping techniques in real time, and evaluate the legitimacy of a URL.

FastAPI backend that provides a real time URL classification API to consume and Streamlit front end with interactive user interface is the architecture of the system. The Streamlit app allows users to input URLs and in turn talks to the FastAPI backend to see if the URL has a high probability of being a phishing scam or not. Feedback mechanism is also present in SafeNet which allows users to identify whether the system's classification is correct or not and it makes the system learn and improvements are made continuously. This is an adaptive learning approach which makes it a safe choice because it is effective in spite of the constant change in the techniques used to emulate phishing emails.

II. LITERATURE SURVEY

In this work, [1] Lamina et al. explore AI power phishing hello detection methods with the machine learning tools to prevent phishing websites and counterfeit emails. Through using the advanced AI algorithms, their study gives us an efficient way to figure out phishing attempts and thereby better prevention and response. As in [2], Prince et al. also highlighted how data-driven AI techniques help to increase the threat identification and in turn, the speed of reaction to a threat. This is actually a very important work that shows that AI can analyze pretty big data in real time, so it can detect and mitigate thousands of cyber threats very quickly.

In the context of cyber threats, Sadaram et al. [3] investigated use of AI for real time detection of cyber threats through use of the machine learning and anomaly detection techniques. Indeed AI can identify abnormal behaviors in the systems in their study and this could help to early detect and mitigate potential threats. Supervised AI classifiers were used to analyze big data environments to defend against phishing and intrusion by Ali [4]. With this in mind, it looked into how AI models can use very large files of data to compare normal and malicious behaviour in order to help secure the workings of large scale systems. Arun and Abosata [5] looked at another generation of phishing attacks, grown in AI empowered browsers. As phishing techniques continue to evolve, AI-driven browsers will likely help develop phishing schemes and so the authors warned that more advanced detection and prevention strategies will be needed. AI can help in the task of detecting and preventing phishing and other cyber attacks was discussed by Naseer [6]. The study looked at how the patterns and behaviors which indicate potential cybersecurity breach can be picked up by the AI.

In the second part, Banu [7] used AI techniques for protecting digital identities specially on the prevention of fraud on online transactions. In an increasingly digital world, digital transactions are increasing rapidly, and AI's ability to verify user identities before committing to an online transaction and preventing fraudulent transactions are of great importance. In the work by Kwaku (8), the challenges and innovations in AI based phishing detection systems were reviewed. His work also highlights the main shortcomings of existing systems and suggests promising ways in which AI is likely to be helpful in bypassing these limitations, enhancing phishing defenses with a high degree of accuracy and reliability.

IoT smart home systems are applied by Fatima et al. [9] with the use of AI to enhance phishing detection and mitigation. Since smart homes are becoming a significant part of everyday life, AI is needed to ensure the domestic IoT devices are safe from phishing and other cyber threats. Real time prediction of vulnerabilities and attack vectors under the study of Malik [10] was examined. The research concentrated on how such predictive analysis of cyber security data by AI algorithms can aid organizations to be proactive in reducing the exposure to attacks. Next generation methods for detecting and prevention of AI phishing scams were investigated by Kumar [11]. In a study, phishing tactics with which AI works are being explored and advanced AI models to differentiate between legitimate and malicious content are being proposed to make detection systems more effective. To address AI's role in detecting threats in cybersecurity infrastructures, Muthusamy [12] was interested in. Machine learning algorithms are detailed in the study as well as the ability of AI powered threat detection systems to use massive

amounts of security data and analyze it in real time for detecting potential security breaches.

Ansarullah et al. [13] analyzed different AI based approaches for antimalware detection and prevention from the advanced malware. One of their contributions was to illustrate the power of AI in improving the old-school malware detection tools through a deep learning and neural network to spot the evolving malware activities and stop the systems from infiltrating. Shabir and Khalid discussed the next level of AI used to detect and assess fraud in financial services. What they found was that the capability of AI to implement analysis of the transaction pattern and use the predictive model to evaluate the risk had significant value in ensuring the safety and integrity of the systems related with finance. Kota [15] suggested a microservices based real time fraud detection system with AI. The institute puts forth this innovative approach of using the modular microservices so that the scalability and flexibility of fraud detection systems within financial institutions are increased, thus reducing the fallout from security breaches, and timely reacting to fraudulent activities. This conclusion further fortifies the growing number of instances in which AI is utilized in the prevention of fraud and the adaptation and output of cybersecurity systems as per the evolving threats.

III. PROPOSED METHODOLOGY

The proposed SafeNet is an AI based phishing detection system which combines a set of machine learning models including one being characterized by feature engineering and real time classification that are able to perform accurate phishing detection and prevention. The methodology uses powerful techniques i.e. PCA (Principal Component Analysis) and RFE (Recursive Feature Elimination) to reduce the dimension of data and select the best features. In this multi model we use Neural Networks (NN), K-Nearest Neighbors (KNN), Naive Bayes, Random Forest and Support Vector Machines (SVM) to have highest detection accuracy and robustness. Web scraping and WHOIS queries are used together with a range of URL-based features like suspicious keywords, web traffic and domain registration details as the system extracts these features dynamically. The methodology consists of supporting a real time prediction through FastAPI API or interactively through a user interface streamed with Streamlit featuring continuous update of user feedback at every step taken.

A. Data Collection and Preprocessing

For the SafeNet phishing detection system, data collection is the collection of datasets which contain legitimate and phishing URLs respectively. These datasets have been created from the publicly available repositories such as OpenPhish and PhishTank and provide a number of labeled URLs. Such datasets provide a number of features such as Domain IP and URL length features, and traffic related data which are important for building a good phishing detection system. Real time threat intelligence feeds are also integrated so the system will be able to detect emerging phishing tactics as they change once something new is just released. Raw data is not clean enough for machine learning models. So, this data preprocessing phase is important to come to a point when raw data is clean enough. In this phase we are going to deal with the missing values, encoding of categorical features and normalization of continuous features to prepare the data for training.

This process can be done only if feature scaling is applied and when models such as KNN, SVM, and random forest are used. First, we apply StandardScaler to normalize the data, so that features are normalized over a standard scale, which facilitates faster and more effective codes against models. Apart from that, dimensionality reduction and feature selection techniques such as PCA (Principal Component Analysis) and RFE (Recursive Feature Elimination) are applied on the given sentence. PCA can help reduce the number of features while keeping important variance and reduce the risk of overfitting due to model performance. Used for recursively removing less important features, the models are trained only with the most important data in this process.

B. Feature Engineering

The SafeNet phishing detection system relies heavily on feature engineering to extract meaningful information from URLs that can improve accuracy of the model. In this phase, a few important features like URL length, Hostname Length, Presence of Special Characters (@, .., /, - etc.), and presence of suspicious keywords as "secure", "login", "account", and "signin" are computed. This is the perfect definition of phishing since fraudulent URLs are full of unusual patterns or keywords that try to mislead visitors into thinking they are visiting a legitimate site. Domain related features, such as domain age and domain registration length also need to be extracted because newly registered domains, or short registration periods usually correspond to phishing websites.

Even more, the feature set includes web traffic data such as Alexa rank. Phishing websites have low web traffic or net traffic and are virtually non-existent, unlike legitimate websites which are well established on the

web. Web scraping and WHOIS queries to dynamically get information about the URL's domain, and its web traffic are used by SafeNet. Real time features, which are important in distinguishing phishing sites from legitimate ones are played. PCA and RFE further integrate feeding the feature set to further reduce noise and omit irrelevant features to feed the models only on most useful information of detecting phishing attempts.

C. AI Model Development

SafeNet employs a multi-model approach: a number of machine learning algorithms are merged in such a way that the combination of them provides robust and accurate phishing detection. By using TensorFlow and Keras frameworks, the Neural Network (NN) model is developed. It is a model of a number of layers of neurons to allow it to learn complex patterns on the feature space. It trains the neural network on the features extracted from the URLs, URL length, domain registration, and suspicious keywords. With the information derived from the training data, the NN model attempts to classify a URL into the phishing or legitimate categories. The model is fine tuning its architecture in such a way that it is more accurate in classifying, also it outperforms other models in characteristic precision and recall of the detection.

Furthermore, K-Nearest Neighbors (KNN), Naive Bayes, Random Forest, and Support Vector Machine (SVM) are used as base models to identify their performance at predicting the label of incoming records. The KNN is a non parametric method characterized as it classifies URLs based on which of their nearby neighbors in feature space have majority, that is, it classifies a URL by predicting the closest class in terms of number ofgetUrl's models. The Naive Bayes model considers Bayes' theorem to estimate the probability a URL is being phished or not. Random Forest is an ensemble of decision trees to predict, while the SVMs select the best hyperplane that can separate phishing and legitimate URLs. The preprocessed and feature engineered data is used to train each model and the performance of the model is evaluated using accuracy, precision, recall, F1 score, etc. These evaluation metrics are used to choose the final model selection which will give highest detection efficiency.

D. Real-Time Detection and API Integration

SafeNet's phishing detection system is a running real time classification of URLs implemented as a combination of a FastAPI backend and a Streamlit front end. When a user pastes in a URL, the FastAPI backend consists of a request which is processed by the backend to extract the relevant features using steps listed in Section A and B. Then, the features are given to the trained machine learning models for the classification. The collective prediction is used to determine whether the URL belongs to a legitimate or a phishing site, and the system guarantees high accuracy by using an ensemble approach. Moreover, the feedback from the users is included as a part of the system through the Streamlil framework app, where the user can give feedback about the accuracy of the prediction. Once the prediction is made, and the prediction is incorrect, the feedback is logged and stored into user_feedback.json to train and retrain continuously on the new phishing tactics for SafeNet to adjust to.

Using the Streamlit front end gives users an easy to use interactive UI to read in URLs and see real time results of classification. Additionally, the system also gives a confidence score of the model that tells how confident the model is on the prediction. The users can share their feedback against the classification which would then be used to retrain the models and improve the accuracy of the system. It's this feedback loop where SafeNet gets to continually improve the phishing detection capabilities using emerging threats as a guide. SafeNet is not only a very good phishing detection system, but it is also scalable and can adapt to the fast growing world of cybersecurity threats using the power of FastAPI and Streamlit.

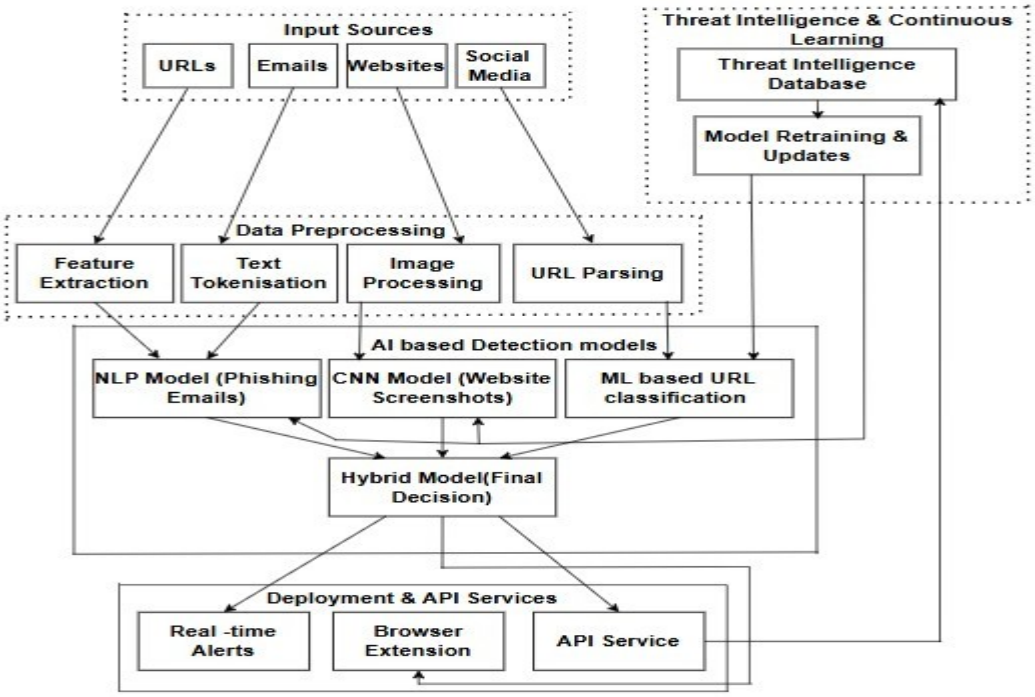

Figure 1 System Architecture

IV. RESULTS AND DISCUSSION

In the following, we show the results achieved by the SafeNet phishing detection system. Different machine learning models such as Neural Networks (NN), K-Nearest Neighbors (KNN), Naive Bayes, Random Forest and Support Vector Machines (SVM) are compared to evaluate the performance of the system. A set of URL-based features such as the domain age, suspicious keywords, web traffic, domain registration details are used to train these models. In addition, we then analyze the dimensionality reduction techniques like PCA (Principal Component Analysis) and RFE (Recursive Feature Elimination) for reducing dimensionality and selecting the most relevant features. Then results are compared based on accuracy, precision, recall and F1 score.

A. Model Performance Evaluation

In the first phase of evaluation, we compared the different models using several of the key metrics and selected the model based on them. The results of accuracy, precision, recall and F1 score for each model are presented in Table 1 below. All the other models achieved the same accuracy between 61.3 to 95.6 %, precision between 56.8 to 96.3 % and recall between 60.4 to 97.3 % but the NN model Table 1 outperformed all other models and achieve the highest accuracy of 97.4 %, precision of 96.8 % and recall of 97.2 %. The results of NN model were compared with Random Forest and SVM models which were competitive but were not close enough to NN in terms of precision and recall. KNN and Naive Bayes performed relatively poorly, especially in recall, which indicated that they were poor at finding phishing URLs.

Table 1: 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. Visualizations

To better understand the relationships between different features and the target variable (status), several visualizations were created. One of the primary visualizations is a boxplot showing the distribution of URL length for legitimate and phishing URLs. Figure 2 below presents this boxplot, which illustrates that phishing Figure 2 URLs tend to have a significantly different distribution of length compared to legitimate URLs. This feature proved to be useful in differentiating phishing from legitimate URLs, as phishing URLs tend to be longer and more complex.

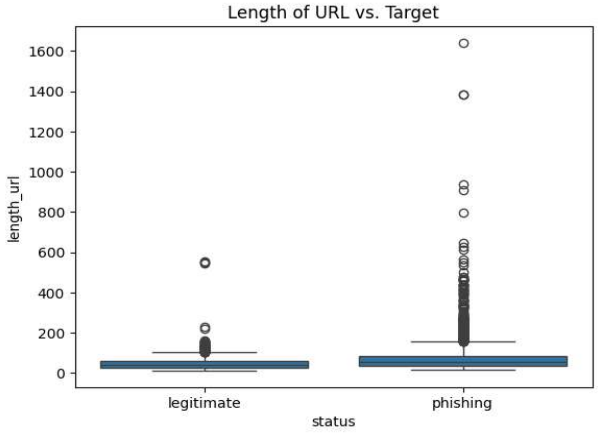

Figure 2: Length of URL vs. Target

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.

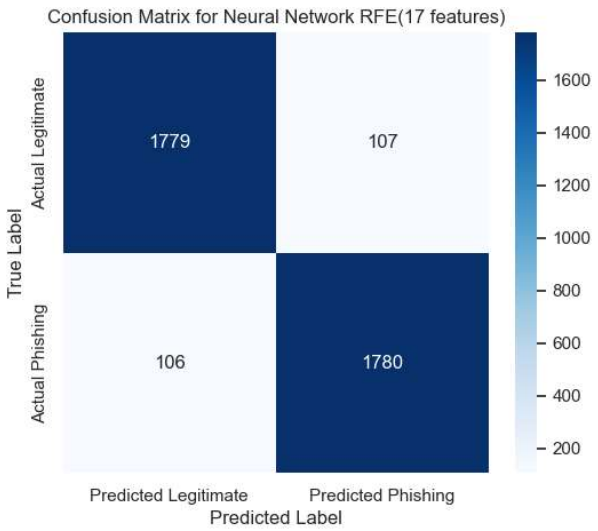

Figure 3: Confusion Matrix

This confusion matrix Figure 3 displays the performance of the NN model with Recursive Feature Elimination method, where the True Positive, True Negative, False Positive, and False Negative rates are visualized for both Legitimate and Phishing URL classifications.

C. False Positive and False Negative Rates

The system also measures the False Positive Rate (FPR) and False Negative Rate (FNR) for each model to assess their ability to correctly classify phishing URLs and avoid misclassifications. The NN model showed the lowest False Positive Rate (FPR) and False Negative Rate (FNR), indicating that it successfully minimized both incorrect classifications. In contrast, Table 2 the Naive Bayes model showed the highest FPR and FNR, which suggests that it was prone to misclassifying phishing URLs as legitimate and vice versa. The Random Forest and SVM models performed well in terms of minimizing FPR and FNR, but they still lagged behind the NN model.

Table 2: 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. Model Tuning and Hyperparameter Optimization

Model tuning and hyperparameter optimization were also critical components in improving the performance of the phishing detection system. Grid Search and Random Search techniques were employed to fine-tune the parameters of models like Random Forest, SVM, and KNN. For example, the Random Forest model was optimized by adjusting the number of trees (n_estimators) and the maximum depth of the trees (max_depth). Similarly, the SVM model was fine-tuned by selecting the optimal C (regularization parameter) and kernel type. These hyperparameter adjustments led to improved performance, particularly in terms of

precision and recall. However, the NN model required fewer adjustments and achieved superior performance without extensive hyperparameter tuning.

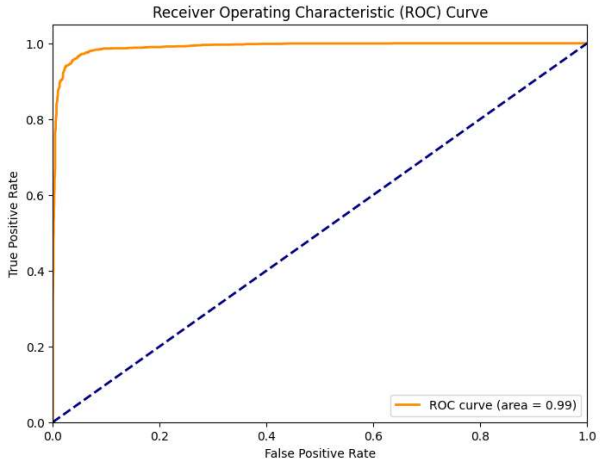

Figure 4: Receiver Operating Characteristic (ROC) Curve

This ROC curve Figure 4 plots the True Positive Rate against the False Positive Rate, highlighting the model's ability to accurately classify phishing URLs. The AUC (Area Under Curve) is 0.99, indicating excellent performance.

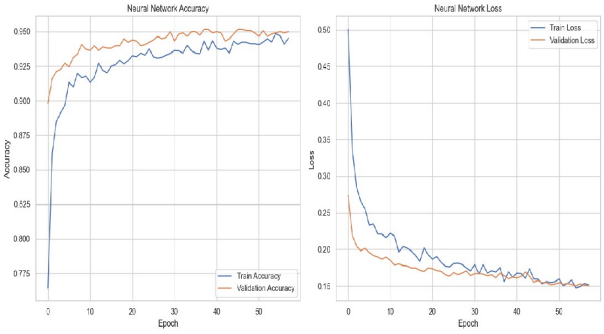

Figure 5: Neural Network Accuracy and Loss Comparison

This plot Figure 5 demonstrates how the incurred accuracy and loss have impact on the model's performance. Both the training and validation accuracies and losses have been illustrated for comparison.

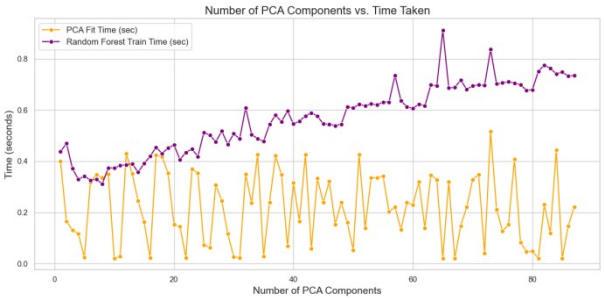

Figure 6: Number of PCA Components vs. Time Taken

This chart Figure 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..

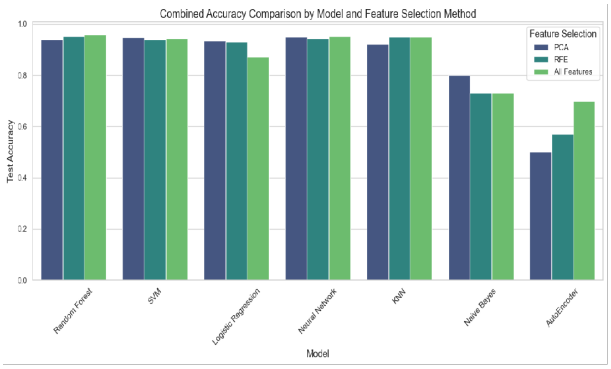

Figure 7: Combined Accuracy Comparison by Model and Feature Selection Method

This bar graph illustrates the accuracy comparisons between the trained models by model and feature selection method.

E. Discussion and Future Improvements

The results from the SafeNet system indicate that the NN model significantly outperforms the other models in terms of detection accuracy, precision, and recall, making it the preferred model for phishing detection. The incorporation of advanced feature engineering techniques such as PCA and RFE contributed to the improved performance by reducing dimensionality and selecting the most significant features. Random Forest and SVM also performed well, though they were slightly less accurate in detecting phishing URLs compared to NN. KNN and Naive Bayes showed lower performance, particularly in recall, which highlights

their limitations in detecting phishing attempts. The integration of user feedback through Streamlit and FastAPI allows the system to continuously learn and adapt to emerging phishing techniques, making it an adaptive and dynamic solution to phishing detection.

Future improvements could involve incorporating Reinforcement Learning (RL) to enhance the system's adaptability to new phishing tactics and further optimize model performance by continuously learning from the feedback and new data.

V. CONCLUSION

In this paper, we describe a system, SafeNet, which employs multiple machine learning models (Neural Networks (NN), K Nearest Neighbors (KNN), Naive Bayes, Random Forest, and Support Vector Machines (SVM)) to detect phishing attacks, except for which one will provide a robust phishing protection. It used PCA (Principal Component Analysis), RFE (Recursive Feature Elimination) and so forth for feature selection and reduced dimensionality, which resulted in better accuracy and performance. Phishing detection result is superior to the NN model compared to other models in accuracy, precision and recall. However, real time detection has been achieved using FastAPI and an interactive interface (Streamlit) which made the system more user friendly and pliable at the same time user feedback is incorporated to constantly improve model performance. Although the current model is effective, future work will be on seeking how the Reinforcement Learning (RL) might help improve system adaptability even more by learning from new phishing tactics and dynamically changing the model's response to new threats. In addition to which, combining blockchain base verification with extending the system into mobile platforms can provide more all-inclusive and scalable phishing protection in this evolving security landscape.

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.