iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84461

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Phishing Website Detection Using A Stacked Hybrid Model With Explainable AI

Adigarla Venkata Naga Mounika[1], Dr. T. Siva Ramakrishna[2]
[1]*Master of Computer Applications, Department of Computer Science and Engineering, University College of Engineering Kakinada (UCEK), JNTU Kakinada, Andhra Pradesh, India*
[2]*Associate Professor, Department of Computer Science and Engineering, University College of Engineering Kakinada (UCEK), JNTU Kakinada, Andhra Pradesh, India*

Abstract: ***Phishing attacks have become a critical cybersecurity challenge due to the increasing use of fraudulent websites and malicious URLs to deceive users and obtain sensitive information. Conventional phishing detection approaches such as blacklist-based filtering and static rule-based techniques often fail to recognize newly generated phishing websites because they depend on previously identified threats. To overcome these limitations, this work presents a Phishing Website Detection System Using a Stacked Hybrid Model With Explainable AI, designed to accurately classify websites as phishing or legitimate while providing transparent decision explanations. The proposed framework utilizes a stacking ensemble architecture that combines multiple learning models, including Artificial Neural Network (ANN), Bagging K-Nearest Neighbors (KNN), and Support Vector Machine (SVM), to identify complex patterns from URL-based and website-related characteristics. Logistic Regression is incorporated as a meta-level classifier to combine the predictions of individual models and generate the final classification result. The system uses extracted lexical, structural, and security-related website features to improve detection reliability and enhance model generalization. Furthermore, Explainable Artificial Intelligence (XAI) is integrated through SHAP to analyze the contribution of individual features and provide meaningful explanations behind each prediction. This improves the transparency of the detection process by showing the factors that influence phishing and legitimate classifications. A user-oriented detection interface is also developed, allowing users to submit website URLs and receive immediate classification results along with feature-based explanations. Performance assessment of the developed hybrid model is carried out using a publicly available phishing website dataset. Experimental results evaluated using Accuracy, Precision, Recall, and F1-score indicates that the stacked hybrid model provides reliable and consistent identification of phishing websites.***

Keywords: ***Phishing Website Detection, Stacked Ensemble Learning, Machine Learning, Deep Learning, Artificial Neural Network, Support Vector Machine, K-Nearest Neighbors, XAI, SHAP, Cybersecurity.***

I. INTRODUCTION

Phishing has emerged as one of the most challenging cybersecurity threats in the digital environment, where attackers create fraudulent websites and deceptive URLs to mislead users and obtain confidential information. These attacks commonly target sensitive resources such as account credentials, financial information, and personal records. With the rapid growth of online services, phishing campaigns have become more sophisticated by generating visually similar and newly registered malicious websites, making accurate detection a difficult task. Conventional phishing detection techniques mainly depend on blacklist databases, predefined security rules, and signature-based analysis. Although these approaches are effective for identifying previously reported malicious websites, they often fail to recognize unknown phishing websites that appear frequently. The dynamic nature of phishing attacks creates a need for intelligent detection mechanisms capable of analyzing website characteristics and identifying suspicious patterns automatically.

The advancement of Artificial Intelligence (AI) and Machine Learning (ML) has significantly improved cybersecurity applications by enabling automated threat identification and classification. Machine learning-based phishing detection methods analyze URL patterns, domain characteristics, and website-related attributes to distinguish between legitimate and malicious websites. Several classification algorithms, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Decision Tree, and Logistic Regression, have been applied for phishing classification. However, individual machine learning models may face the limitations in handling complex patterns and limitations in handling complex patterns and achieving consistent detection performance across diverse phishing scenarios.

Deep Learning techniques have further enhanced the capability of cybersecurity systems by learning complex relationships from large-scale feature representations. Artificial Neural Networks (ANNs) are widely used for classification tasks because of their ability to capture nonlinear patterns within data. Despite their effectiveness, single deep learning models may still produce inaccurate predictions when dealing with evolving phishing strategies and highly variable website characteristics. To improve detection reliability, ensemble learning approaches have gained attention by combining multiple learning models to utilize their individual strengths. Stacking-based ensemble learning provides an effective mechanism where predictions from multiple base models are combined through a secondary learning model to generate the final decision. This approach improves generalization capability and reduces dependency on a single classifier. This research presents a stacked hybrid learning framework designed to improve phishing website classification performance. The proposed framework integrates Artificial Neural Network (ANN), Bagging K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) as base learners. The outputs generated by these models are combined using Logistic Regression as a meta-learning model to produce the final classification result. By combining multiple learning approaches, the framework aims to improve phishing detection performance and provide a more reliable solution compared with individual classification techniques. Furthermore, Explainable Artificial Intelligence (XAI) is incorporated into the proposed system to improve transparency and interpretability. SHAP is used to determine the important factors influencing the model decision and provide meaningful explanations behind phishing or legitimate predictions. This helps users understand why a particular website has been classified as malicious or safe. Additionally, a real-time interface is developed to allow users to submit website URLs and obtain immediate detection results along with explanation details. Unlike traditional detection approaches that only provide classification outcomes, the proposed system focuses on combining accurate prediction, model transparency, and practical usability.

Motivated by the limitations of existing phishing detection methods, this work presents a Stacked Hybrid Phishing Website Detection Framework with Explainable AI. The proposed approach aims to provide an effective, interpretable, and scalable cybersecurity solution for identifying modern phishing websites and improving protection against online threats.

II. RELATED WORK

A. *Machine Learning Based Phishing Website Detection*

The rapid growth of phishing attacks has encouraged researchers to develop intelligent detection techniques capable of distinguishing malicious websites from genuine ones. Machine learning has become one of the most widely adopted solutions because it can automatically recognize hidden relationships within website and URL features. Early investigations demonstrated that learning-based models outperform traditional rule-based approaches by adapting to evolving phishing patterns. Early investigations demonstrated that learning-based models outperform traditional rule-based approaches by adapting to evolving phishing patterns. Abu-Nimeh et al. **[1]** carried out a comparative investigation of several machine learning classifiers for phishing website identification. Their experimental study showed that supervised learning algorithms can successfully differentiate deceptive websites from legitimate ones by analyzing extracted website features. Abutaha et al. **[2]** introduced a phishing detection approach centred on lexical analysis of URLs. Their approach examined various URL characteristics such as domain structure, length, special symbols, and other URL-related patterns to differentiate phishing and legitimate websites. The study showed that URL feature analysis plays an important role in detecting suspicious websites. Shahrivari et al. [3] investigated different machine learning approaches for phishing website classification and analyzed the impact of feature selection and model performance. Their research highlighted that appropriate feature representation and optimized classification methods are essential for improving phishing detection accuracy. Dadvandipour and Ganie [4] explored machine learning approaches for identifying phishing-related activities and demonstrated the effectiveness of predictive models in recognizing malicious patterns. Their work emphasized the capability of machine learning techniques in improving automated cyber threat detection. Aiyanyo et al. [5] presented a systematic review of machine learning applications in cybersecurity. Their study discussed how intelligent learning algorithms are increasingly used for detecting cyber threats and improving security mechanisms compared with conventional rule-based approaches.

B. *Deep Learning Approaches for Phishing Detection*

The emergence of deep learning has significantly strengthened cybersecurity applications by enabling automatic extraction of complex feature representations. Unlike conventional machine learning algorithms that rely heavily on manually engineered features, deep neural networks are capable of learning hierarchical patterns directly from large-scale datasets. Sarker [6] provided a detailed overview of deep learning-based cybersecurity solutions and explained how neural network models can improve threat identification by learning hidden relationships within security data.

Mughaid et al. [7] introduced a deep learning framework for phishing website detection. Their approach utilized neural network-based learning methods to analyze phishing-related features and achieved improved classification performance compared with traditional approaches.

Alsubaei et al. [8] introduced a hybrid deep learning framework for phishing detection and cybercrime investigation. The research demonstrated that combining multiple learning strategies can improve detection reliability and reduce incorrect predictions in complex cyber environments. Ozcan et al. [9] introduced a hybrid DNN-LSTM based phishing URL detection approach. Their approach utilizes deep neural networks with sequential learning models can improve the ability to identify complex URL patterns associated with phishing attacks. Newaz et al. [10] presented a framework for accurate phishing website identification using intelligent learning techniques. Their research focused on improving automated detection capability by analyzing website attributes and classification behavior.

C. Hybrid Machine Learning and Ensemble Learning Approaches

Individual Although individual classifiers often achieve satisfactory performance, relying on a single learning algorithm may reduce robustness when confronted with continuously changing phishing techniques. Ensemble and hybrid learning strategies address this limitation by combining the complementary strengths of multiple predictive models. Karim et al. [11] proposed a phishing detection framework using hybrid machine learning techniques using URL-derived features. Their approach combined multiple learning methods to improve classification accuracy and demonstrated that hybrid models provide stronger generalization capability compared with individual classifiers. Helali [12] investigated phishing detection using hybrid machine learning algorithms. The study showed that combining different models can improve prediction reliability by reducing the weaknesses of individual classifiers. Kocyigit et al. [13] introduced an enhanced phishing URL detection approach using genetic algorithm-based feature selection. Their work focused on selecting important website features to improve machine learning efficiency and reduce unnecessary computation. Singh et al. [14] proposed a URL-based phishing prevention system using machine learning methods. The research analyzed URL characteristics and applied classification techniques to identify malicious websites effectively. Raza et al. [15] explored machine learning-based attack detection using optimized feature-based learning approaches. Their demonstrated that feature optimization and advanced classification strategies can improve cybersecurity prediction performance.

D. Explainable Artificial Intelligence in Cybersecurity

Recent developments in cybersecurity have demonstrated that achieving high prediction accuracy alone is insufficient for practical security applications. Modern machine learning models often generate reliable classification results; however, the absence of interpretability makes it difficult to understand the reasoning behind their decisions. Recent studies have emphasized that transparent prediction mechanisms improve the reliability of intelligent security systems and support better decision-making in threat detection environments. systems for improving trust and reliability in cybersecurity solutions. Researchers have increasingly explored Explainable Artificial Intelligence (XAI) as a solution for improving the interpretability of machine learning models used in cybersecurity. Instead of providing only a final prediction, XAI techniques identify the influence of individual input features on the model's decision. Such explanations enable security analysts to verify prediction outcomes, understand feature importance, and analyze the behaviour of intelligent detection systems more effectively. The growing adoption of explainable models has also contributed to improving confidence in automated cybersecurity applications. In phishing website detection, explainability has become an important component for understanding why a website is classified as suspicious or legitimate. Techniques such as SHAP provide both overall feature importance and instance-level explanations, allowing users to examine the contribution of URL characteristics and website attributes to each prediction.

E. Research Gap and Proposed Approach

Numerous studies have reported encouraging results by applying machine learning and deep learning techniques to phishing website detection. Despite these advancements, many existing methods rely on a single classification algorithm, which may not consistently maintain high performance when confronted with newly emerging phishing strategies. Furthermore, several approaches mainly focus on prediction accuracy while providing limited explanation regarding the factors responsible for classification decisions. To overcome these limitations, this research proposes a Phishing Website Detection Using A Stacked Hybrid Model With Explainable AI. The proposed framework utilizes a stacking ensemble architecture combining ANN, Bagging KNN, and SVM models. Logistic Regression is employed as a meta-classifier to combine individual model outputs and generate the final phishing or legitimate prediction.

Additionally, SHAP-based Explainable Artificial Intelligence is integrated to identify the important features contributing to the prediction process. This improves model transparency, strengthen user confidence, and support practical deployment in cybersecurity applications.

III. METHODOLOGY

The proposed Phishing Website Detection Using a Stacked Hybrid Model with Explainable Artificial Intelligence consists of six major phases, as illustrated in Fig. 1. The framework combines comprehensive data preprocessing, multiple machine learning algorithms, stacking ensemble learning, performance evaluation, and SHAP-based Explainable AI to accurately classify websites as phishing or legitimate while providing interpretable prediction explanations.

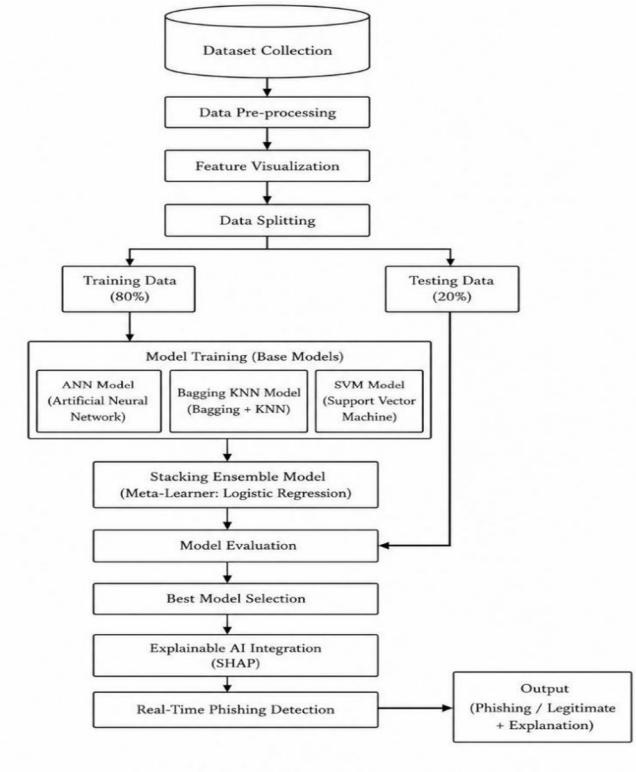


Fig 1. Workflow of the proposed system

A. Dataset Collection and Data Preprocessing

The initial stage involves acquiring a phishing website dataset [16] containing numerous URL-based, domain-related, and webpage behavioral attributes that describe the characteristics of both legitimate and malicious websites.

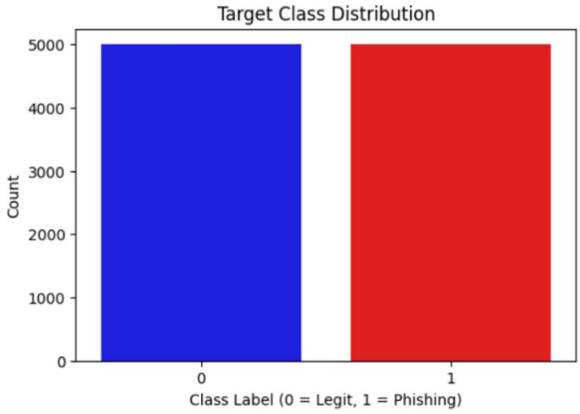


Fig 2. Target Class Distribution

Every instance in the dataset is assigned one of two class labels, where 0 denotes a legitimate website and 1 represents a phishing website.

As illustrated in Fig. 2, the dataset exhibits an almost equal distribution between both classes, with 5,000 legitimate samples and 5,000 phishing samples. A balanced dataset minimizes class imbalance issues and enables the learning algorithms to capture representative patterns from both categories without favoring one class over the other.

Before model development, preprocessing operations are performed to improve data quality. Missing attribute values are identified and handled through appropriate imputation techniques wherever necessary. Duplicate observations are eliminated to prevent redundant information from influencing model learning. Furthermore, invalid records and inconsistent feature values are inspected and corrected to ensure dataset reliability. To improve the effectiveness of the learning algorithms, numerical attributes are standardized using the StandardScaler technique. This transformation converts each feature into a standardized distribution with zero mean and unit variance. Feature standardization prevents attributes with larger numerical values from dominating the optimization process and significantly improves the convergence behaviour of models such as ANN and SVM. These preprocessing activities enhance dataset consistency, eliminate unnecessary noise, and generate a clean and standardized dataset suitable for efficient machine learning.

B. Feature Visualization and Exploratory Data Analysis

Once the preprocessing stage is completed, exploratory analysis is carried out to examine the statistical properties of the dataset before training the prediction models. This phase aims to understand the distribution and significance of different phishing-related attributes. Initially, the target class distribution is analyzed to verify the balance between phishing and legitimate website samples. Additional visualization methods are employed to study the behavior of individual features, detect abnormal observations, and understand variations across website characteristics.

Feature exploration provides valuable insights into the underlying data structure and assists in identifying meaningful attributes that contribute to phishing detection. This preliminary investigation ensures that the dataset is suitable for subsequent machine learning analysis and supports the development of robust classification models.

C. Data Splitting and Base Model Training

The processed dataset is partitioned into training (80%) and testing (20%) subsets. The training portion is utilized to learn the underlying patterns associated with website characteristics, whereas the testing portion is reserved exclusively for measuring the predictive performance of the trained models using previously unseen data. Three independent machine learning algorithms are trained using the training subset before their outputs are integrated through the proposed ensemble strategy. The models are:

1) Artificial Neural Network (ANN): The Artificial Neural Network is designed to learn complex nonlinear relationships among phishing website features through interconnected computational neurons arranged in multiple layers. The architecture consists of an input layer, several hidden layers, and a binary output layer responsible for classification. To improve learning stability and reduce overfitting, Batch Normalization and Dropout layers are incorporated into the network architecture. During training, the ANN automatically discovers meaningful feature representations and produces prediction probabilities indicating whether a website belongs to the phishing or legitimate class.

2) Bagging K-Nearest Neighbors (Bagging KNN): The Bagging KNN classifier combines the Bootstrap Aggregating (Bagging) technique with the conventional K-Nearest Neighbors algorithm. Multiple bootstrap datasets are generated from the original training data, and an individual KNN classifier is trained on each subset independently. During prediction, the outputs from all trained classifiers are combined using majority voting to produce the final decision. This ensemble strategy decreases prediction variance, increases robustness against noisy samples, and provides better generalization than a standalone KNN classifier.

3) Support Vector Machine (SVM): Support Vector Machine (SVM) is utilized to construct an optimal separating boundary between different classes within the feature space. The algorithm focuses on selecting support vectors that define the decision boundary while maximizing the classification margin. Because phishing datasets generally contain high-dimensional feature spaces, SVM is well suited for identifying discriminative boundaries and maintaining reliable prediction performance across unseen data. The prediction probabilities generated by these three individual classifiers are subsequently forwarded to the stacking ensemble model.

D. Stacking Ensemble Learning

The fourth stage develops the proposed hybrid classification framework using a stacking ensemble strategy. Instead of relying on the prediction of a single machine learning algorithm, the proposed system integrates the strengths of multiple classifiers. The outputs produced by the ANN, Bagging KNN, and SVM models serve as input features for a Logistic Regression meta-classifier. This second-level learning model analyzes the prediction behaviour of each base classifier and determines the optimal combination required to generate the final website classification. By exploiting the complementary learning capabilities of different machine learning algorithms, the stacking ensemble minimizes the weaknesses of individual classifiers while improving classification reliability, prediction stability, and overall phishing detection accuracy.

E. Model Evaluation

The effectiveness of the developed classifiers is assessed using the reserved testing dataset. The predictive performance of Logistic Regression, Bagging K-Nearest Neighbors, Support Vector Machine, Artificial Neural Network, and the proposed Stacking Hybrid Model is compared using multiple evaluation metrics. Experimental observations indicate that the stacking framework consistently delivers superior performance by effectively combining the strengths of the individual classifiers. The comparative evaluation demonstrates that ensemble learning significantly improves phishing detection capability while reducing classification errors.

F. Explainable AI Integration and Real-Time Phishing Detection

To enhance the interpretability of the proposed phishing website detection framework, SHAP is incorporated as the explainability technique. Rather than providing only the prediction result. SHAP quantifies the contribution of each website feature to the model's decision by assigning an importance score derived from cooperative game theory. This enables the framework to present meaningful explanations alongside every classification outcome. The explainability process produces both global and local interpretations of the trained model. Global explanations reveal the features that have the greatest overall influence across the complete dataset, whereas local explanations describe how individual feature values affect the prediction for a specific website instance. These visual and numerical explanations allow users to understand the reasoning behind the model's decisions instead of treating it as a black-box system. The trained stacking model, together with SHAP explanations, is deployed within a real-time phishing detection system. Users provide a website URL, from which relevant features are automatically extracted. Along with the final classification, the system presents feature-level explanations that indicate the factors responsible for the decision. This combination of accurate prediction and transparent interpretation strengthens user confidence, and improves the practical usability of the proposed framework.

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The effectiveness of the proposed phishing website detection framework is validated through a series of experiments involving both individual machine learning models and the proposed stacking ensemble model. Performance comparisons are carried out to examine the predictive capability of each classifier using several widely adopted evaluation measures. These experiments provide a comprehensive assessment of the proposed approach and demonstrate its effectiveness in identifying phishing websites. To ensure a fair comparison, the performance of all classifiers is evaluated using common classification metrics. These metrics provide a standardized basis for comparing the individual classifiers with the proposed hybrid stacking model and help identify the most effective approach for the classification task. A brief description of each evaluation metric is presented below.

1) Accuracy

Accuracy represents the proportion of correctly classified samples among the total number of observations processed by the model.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2) Precision

Precision measures the proportion of predicted positive instances that are actually positive.

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

3) Recall

Recall measures the ability of a model to correctly identify all actual positive instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

4) F1-Score

The F1-Score combines Precision and Recall into a single performance measure by calculating their harmonic mean.

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

B. Performance Comparison of Models

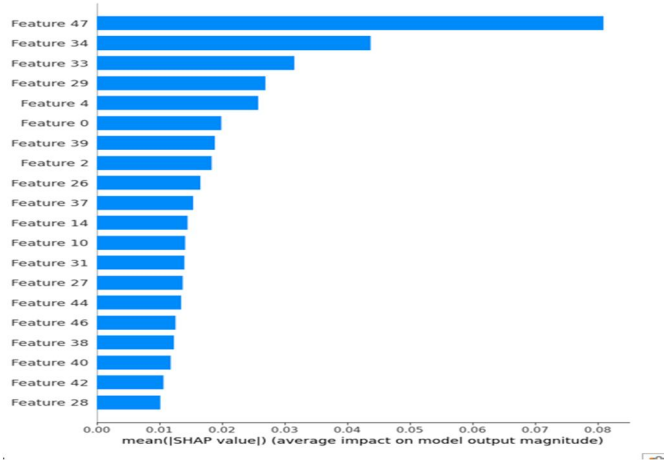
Table 1 presents the comparative performance evaluation of all implemented models. The comparison highlights the effectiveness of both individual classifiers and the proposed hybrid stacking framework.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	95.20%	94.48%	96.00%	95.23%
Bagging KNN	96.00%	95.63%	96.40%	96.01%
SVM	96.90%	96.16%	97.70%	96.92%
ANN	97.90%	97.42%	98.40%	97.91%
Hybrid Stacking Model	98.05%	97.71%	98.40%	98.06%

Table-1 performance comparison of models

The experimental results demonstrate that all models achieved effective phishing classification capability. Logistic Regression provided a strong baseline performance; however, advanced learning approaches achieved better results. The Bagging KNN model improved performance through ensemble-based learning, while SVM achieved higher accuracy by constructing optimized classification boundaries. ANN further enhanced detection performance by automatically learning nonlinear relationships between URL characteristics. The proposed Hybrid model delivers the highest performance with 98.05% accuracy and 98.05% F1-score, proving its effectiveness in combining multiple learning strategies. The improved results indicate that ensemble-based integration provides better generalization capability and reduces classification errors in phishing website detection.

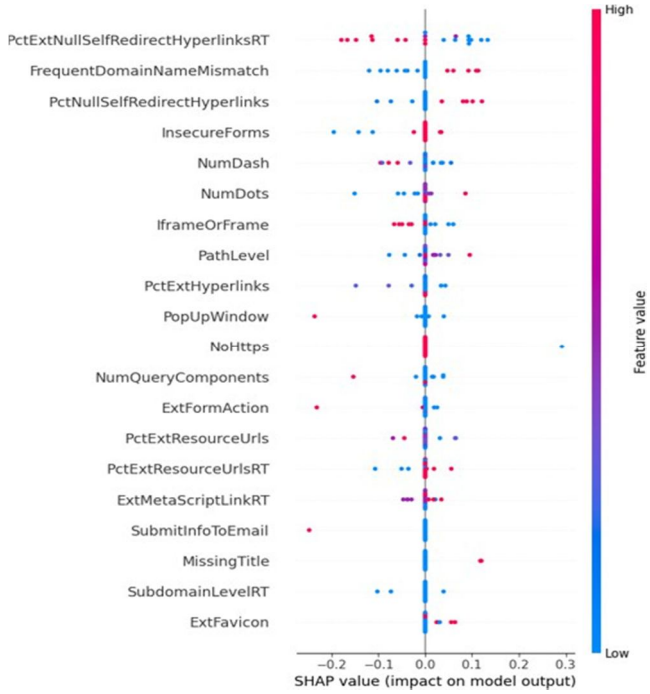
C. Explainable AI Analysis Using SHAP



(a) SHAP Feature Importance Bar Plot

Fig (a) presents the SHAP feature importance bar plot, which ranks the input features according to their average contribution to the predictions generated by the proposed hybrid stacking model. Features with higher SHAP values exert a greater effect on the final classification generated by the model. The analysis reveals that Feature 47 has the highest impact on classification, followed by Feature 34, Feature 33, Feature 29, Feature 4, Feature 0, Feature 39, and Feature 2. These features contribute more significantly than the remaining attributes when distinguishing between different website classes. On the other hand, features with smaller SHAP values have a comparatively lower influence on the final prediction.

The feature importance analysis demonstrates that the proposed model primarily relies on a subset of highly informative attributes instead of assigning equal importance to every input feature. This selective learning behaviour enables the hybrid classifier to make reliable predictions while improving computational efficiency and enhancing the interpretability of the overall detection framework.



(b) SHAP Summary Plot

Fig (b) illustrates the SHAP summary plot, offering a comprehensive view of how different feature values affect the prediction behaviour of the proposed stacked model. Each point in the figure represents the SHAP value of a single observation, while the colour scale indicates whether the corresponding feature value is relatively high or low. The summary plot indicates that attributes such as PctExtNullSelfRedirectHyperlinksRT,FrequentDomainNameMismatch, PctNullSelfRedirectHyperlinks, InsecureForms, NumDash, NumDots, IframeOrFrame, PathLevel, and PctExtHyperlinks exert considerable influence on the classification process. Depending on their feature values, these attributes either increase or decrease the model's prediction score, demonstrating that each feature contributes differently to the final decision.

Unlike the feature importance bar chart, which only presents the overall ranking of attributes, the SHAP summary plot presents both the magnitude and direction of feature influence. This enables a deeper understanding of the model's prediction behaviour and provides transparency by explaining how individual website characteristics affect each classification result. Consequently, the integration of SHAP enhances the interpretability of the proposed phishing website detection system and increases user confidence in its real-world deployment.

D. Discussion

The experimental findings confirm that the proposed stacked ensemble framework performs efficiently in identifying phishing websites. Each base classifier, including Bagging KNN, Support Vector Machine, and Artificial Neural Network, successfully learned discriminative patterns from URL and website-related features, while Logistic Regression provided a reliable baseline for comparison. Although each model achieved strong classification performance, the proposed Hybrid Stacking model delivered superior results by combining the predictions of the base classifiers through a Logistic Regression meta-learner.

The ensemble learning strategy minimized prediction errors while improving the overall evaluation metrics. In addition, SHAP-based explainability increased the transparency of the prediction process by identifying the contribution of individual features Consequently, the proposed framework offers both high predictive capability and clear decision explanations, making it well suited for practical cybersecurity applications.

V. CONCLUSION AND FUTURE SCOPE

A. Conclusion

The proposed machine learning-based phishing website detection framework was successfully developed to identify malicious and legitimate websites using URL-based feature analysis and intelligent classification techniques. The architecture combines Artificial Neural Network, Bagging K-Nearest Neighbors, Support Vector Machine, and Logistic Regression within a stacking ensemble framework. This integration enhances the reliability and effectiveness of phishing website classification. The complete workflow includes dataset preparation, preprocessing, exploratory analysis, model construction, performance assessment, and explainable prediction analysis. Experimental evaluation showed that the implemented machine learning models are capable of identifying hidden patterns within URL characteristics and effectively distinguishing phishing websites from legitimate websites.

Among all evaluated approaches, the stacking ensemble consistently produced the strongest predictive performance with an accuracy of 98.05%, precision of 97.71%, recall of 98.40%, and F1-score of 98.05%. The results obtained from confusion matrix analysis confirm that the Hybrid approach produces fewer false classifications compared with individual machine learning models. This indicates that combining multiple learning approaches improves classification reliability and enhances phishing detection performance. One of the primary contributions of this research is the incorporation of Explainable Artificial Intelligence (XAI) through the SHAP interpretation technique. SHAP enables detailed interpretation of the trained model by measuring how individual URL-based features influence the final prediction, thereby improving the transparency of the classification process. Features such as suspicious redirects, domain-related characteristics, insecure forms, and abnormal URL structures were observed as significant contributors toward phishing classification. The proposed framework not only provides accurate phishing detection but also improves transparency and trust by explaining why a particular website is classified as malicious or legitimate. The combination of machine learning-based classification and explainable prediction analysis makes the developed system suitable for real-world cybersecurity applications such as browser protection, automated threat detection, and online security monitoring. Overall, the proposed phishing detection framework demonstrates that intelligent machine learning techniques combined with explainability methods can provide an efficient, reliable, and interpretable solution for detecting modern phishing attacks.

B. Future Scope

Future research may concentrate on designing a more adaptive phishing detection framework that can effectively recognize newly emerging phishing attempts under real-time operating conditions. Since attackers continuously modify URL structures, webpage designs, and domain characteristics, future models can incorporate online learning techniques that allow the system to update its knowledge continuously without complete retraining.

The proposed framework may also be enhanced by incorporating advanced deep learning architectures, including transformer networks, graph neural networks, and other hybrid learning models capable of extracting sophisticated phishing patterns from large cybersecurity datasets. These approaches may improve adaptability against newly generated phishing websites and zero-day attacks.

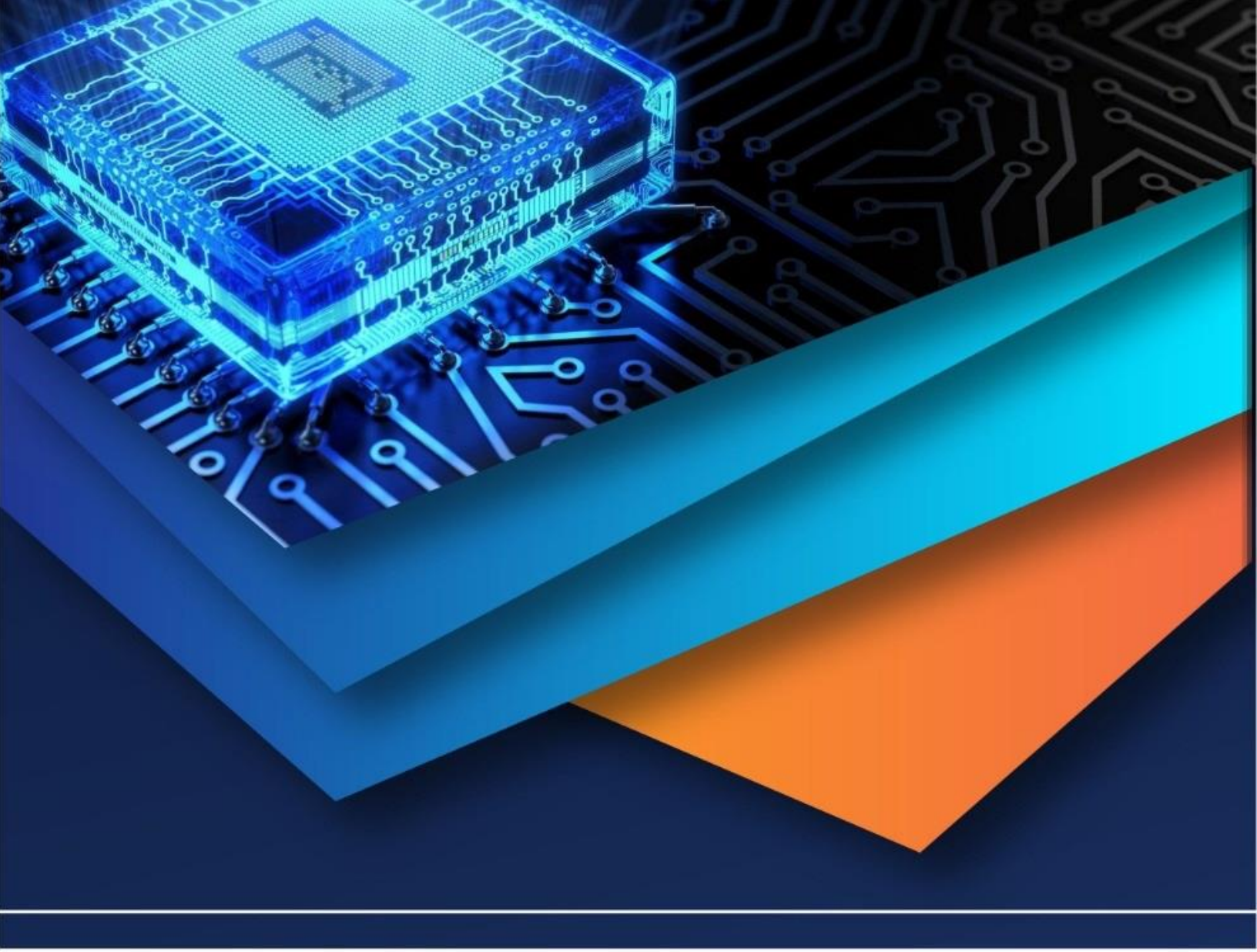
Another possible extension is the development of a real-time phishing detection platform integrated with web browsers, mobile applications, and cybersecurity monitoring systems. Deployment on cloud and edge computing environments can enable faster prediction and continuous protection for users during online activities.

Furthermore, the explainability component can be enhanced by combining SHAP with other interpretation techniques to provide more detailed security reports and human-understandable explanations. A user-friendly dashboard can also be developed to visualize detected threats, important URL characteristics, risk levels, and model decisions.

REFERENCES

[1] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, “A Comparison of Machine Learning Techniques for Phishing Detection,” eCrime Researchers Summit, 2007.
[2] M. Abutaha, M. Ababneh, K. Mahmoud, and S. A.-H. Baddar, “URL Phishing Detection Using Machine Learning Techniques Based on URLs Lexical Analysis,” ICICS, 2021.
[3] V. Shahrivari, M. M. Darabi, and M. Izadi, “Phishing Detection Using Machine Learning Techniques,” arXiv, 2020.
[4] S. Dadvandipour and A. G. Ganie, “Analyzing and Predicting Spear Phishing Using Machine Learning Methods,” 2020.
[5] I. D. Aiyanyo, H. Samuel, and H. Lim, “A Systematic Review of Defensive and Offensive Cybersecurity with Machine Learning,” Applied Sciences, 2020.

[6] I. H. Sarker, "Deep Cybersecurity: A Comprehensive Overview from Neural Network and Deep Learning Perspective," SN Computer Science, 2021.
[7] S. Mughaid et al., "An Intelligent Cyber Security Phishing Detection System Using Deep Learning Techniques," Cluster Computing, 2022.
[8] F. S. Alsubaei et al., "Enhancing Phishing Detection: A Novel Hybrid Deep Learning Framework for Cybercrime Forensics," IEEE Access, 2024.
[9] A. Ozcan et al., "A Hybrid DNN-LSTM Model for Detecting Phishing URLs," Neural Computing and Applications, 2023.
[10] A. Newaz et al., "A Sophisticated Framework for the Accurate Detection of Phishing Websites," arXiv, 2024.
[11] A. Karim et al., "Phishing Detection System Through Hybrid Machine Learning Based on URL," IEEE Access, 2023.
[12] R. G. M. Helali, "Phishing Detection Using Hybrid Machine Learning Techniques," 2024.
[13] E. Kocyigit et al., "Enhanced Feature Selection Using Genetic Algorithm for Machine-Learning-Based Phishing URL Detection," Applied Sciences, 2024.
[14] N. T. Singh et al., "An Innovative URL-Based System Approach with ML Based Prevention," IEEE IC2PCT, 2024.
[15] A. Raza et al., "Novel Class Probability Features for Optimizing Network Attack Detection with Machine Learning," IEEE Access, 2023.
[16] S. Tiwari, Phishing Dataset for Machine Learning, Kaggle.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY