International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 22 No. 2(1) (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

ENHANCING PHISHING DETECTION: MACHINE LEARNING APPROACH WITH FEATURE SELECTION AND DEEP LEARNING MODELS

[1] B Rajasri, [2] D Sainath Reddy, [3] B Rushikesh, [4] E Ravi Teja, [5] B Mukhesh
[1]AssistantProfessor, [2345]Students
Department of AIML
Siddhartha Institute of Technology & Sciences, Narapally
rajasribomma@siddhartha.org.in, 24tq1a6629@siddhartha.co.in, 24tq1a6618@siddhartha.co.in, 24tq1a6633@siddhartha.co.in, 24tq1a6616@siddhartha.co.in,

Abstract

Phishing attacks have become a major cybersecurity threat, exploiting users through malicious URLs and deceptive websites. This project proposes a hybrid machine learning framework for phishing URL detection that combines the strengths of XGBoost and a Deep Neural Network (DNN). The system begins with data preprocessing, including feature extraction, scaling, and selection to ensure high-quality input. The processed data is then split into training and testing sets. The dataset used in this project is the Phishing Website Dataset (e.g.,https://www.kaggle.com/datasets/mdsultanulislamovi/phishingwebsite-detection-datasets), which contains labeled instances of legitimate and phishing URLs. XGBoost captures complex patterns in structured data, while the DNN, implemented using PyTorch, learns deep feature representations with dropout to prevent overfitting. The predictions from both models are combined using a hybrid approach to enhance accuracy and robustness. Performance is evaluated using metrics such as accuracy, confusion matrix, and loss curves. Experimental results demonstrate that the hybrid model achieves an accuracy of approximately 96–98%, outperforming individual models and providing reliable and efficient phishing detection. This approach contributes to improved cybersecurity by enabling accurate identification of malicious URLs and enhancing protection against phishing attacks.

Keywords:
Phishing Detection, Machine Learning, Deep Learning, XGBoost, Deep Neural Network, Hybrid Model, Cybersecurity, URL Classification

I. Introduction

In the digital era, the rapid growth of the internet has significantly transformed how individuals and organizations communicate, conduct business, and access information. However, this expansion has also introduced various cybersecurity threats, among which phishing attacks remain one of the most prevalent and dangerous. Phishing is a type of cyberattack where attackers attempt to deceive users into revealing sensitive information such as usernames, passwords, credit card details, and other personal data by impersonating legitimate entities. These attacks are commonly carried out through malicious URLs, fake websites, and deceptive emails, making them difficult to detect using traditional security mechanisms.

Over the years, phishing techniques have evolved in complexity and sophistication. Attackers now use advanced obfuscation strategies, domain spoofing, and social engineering tactics to make malicious URLs appear legitimate. As a result, conventional rule-based and blacklist-based detection systems are no longer sufficient to combat modern phishing attacks. These traditional approaches often fail to identify newly generated or previously unseen phishing URLs, also known as zero-day attacks, which pose a significant risk to users and organizations.

To address these challenges, researchers have increasingly turned to machine learning (ML) and deep learning (DL) techniques for phishing detection. Machine learning models can automatically learn patterns from data and classify URLs as phishing or legitimate based on extracted features. Algorithms such as Support Vector Machines (SVM), Decision Trees, and Random Forest (RF) have been widely used in phishing detection tasks. Among these, ensemble methods like Random Forest and boosting algorithms such as XGBoost have shown strong performance due to their ability to handle structured data and capture complex relationships between features.

Despite the effectiveness of traditional ML techniques, they often rely heavily on manual feature engineering and may struggle to capture deeper and more abstract patterns in data. This limitation has led to the adoption of deep learning models, particularly Deep Neural Networks (DNNs), which can automatically learn hierarchical feature representations. DNNs consist of multiple layers of interconnected neurons that process input data and extract meaningful patterns through non-linear transformations. By using activation functions such as ReLU and regularization techniques like dropout, DNNs can effectively model complex relationships and reduce overfitting.

II. Literature Survey

Kumar et al.[1], 2023 – Detection and Classification of Phishing Websites Using Machine Learning Approach Taking Advantage of Metaheuristic Algorithms Efficiency. This study uses a structured phishing dataset with URL and webpage features such as URL length, IP presence, and subdomains. It applies Voting and Random Forest classifiers optimized with MOA and TWO algorithms, achieving 99% accuracy. The work highlights optimization and feature selection, which relates to our project that further enhances performance using deep learning models like DNN and Tab Transformer.

Sharma et al.[2], 2022 – An Ensemble Learning Approach for Detecting Phishing Websites Using an Entropy-Based Feature Selection Method. This study uses a benchmark dataset of 11,430 instances with 87 features, reduced to 30 using entropy-based information gain. Multiple ML models and a stacked ensemble are applied, achieving 96.597% accuracy. It shows feature selection improves efficiency without losing performance. This relates to our project by emphasizing feature selection and ensemble learning, while we extend it using deep learning models like DNN and TabTransformer.

Patel et al.[3], 2023 – Optimizing URL-Based Phishing Detection Using XGBoost and Relief Feature Selection. This study uses a dataset of 10,000 instances with 50 URL-based features, applying normalization and feature selection techniques like Information Gain, Chi-square, and ReliefF. XGBoost achieved the highest accuracy of 98.8%, while ReliefF reduced features effectively without performance loss. This work relates to our project by emphasizing feature selection and high-performance

models, which we extend using deep learning approaches like DNN and TabTransformer.

Catal et al.[4], (2022) provide a comprehensive review of deep learning techniques for phishing detection across multiple datasets. The study highlights how deep learning models can capture complex patterns and subtle features more effectively than traditional machine learning methods. Their findings show that DL consistently outperforms ML in challenging scenarios, especially with large and diverse data, thereby strongly supporting the use of deep learning in advanced phishing detection systems.

Shahrivari et al.[5],(2020) explore phishing detection using machine learning techniques on both email and URL datasets. The study compares algorithms such as Support Vector Machine (SVM) and Random Forest (RF) to evaluate their effectiveness. Results show that Random Forest outperforms SVM in terms of accuracy and reliability. This work highlights the importance of model comparison and supports the selection of robust ML algorithms in phishing detection systems.

Salahdine et al.[6],(2022) investigate phishing detection using Artificial Neural Networks (ANN) on a dataset of over 4000 email samples. The study demonstrates that ANN models can effectively learn patterns in phishing emails and achieve high accuracy in classification. Their results highlight the strength of neural networks in handling complex data, supporting the adoption of deep learning approaches for improving phishing detection performance.

Patil et al.[7],(2022) present a machine learning approach for phishing website detection using multiple datasets to ensure robustness and reliability. The study compares various ML algorithms to evaluate their performance in identifying phishing websites. By analyzing different models, the research establishes strong baseline methods and highlights the effectiveness of traditional machine learning techniques in detecting phishing attacks, supporting their role in building reliable detection systems.

Lokesh and Gowda.[8],(2020) study phishing detection using machine learning techniques on a web-based dataset. Their research focuses on evaluating different ML models, with Random Forest (RF) achieving the highest accuracy among the tested algorithms. The results demonstrate the effectiveness of RF in identifying phishing websites and highlight the importance of machine learning models in building reliable and accurate phishing detection systems.

Kuraku and Kalla.[9],(2023) propose a phishing URL detection approach that integrates Natural Language Processing (NLP) with machine learning techniques. Using the PhishTank dataset, their method extracts textual and structural features from URLs to improve detection accuracy. By combining NLP and ML, the study demonstrates the effectiveness of hybrid feature engineering, highlighting how such approaches can enhance performance in identifying phishing attacks.

Divakaran and Oest.[10] (2022) examine phishing detection using both machine learning and deep learning techniques on large-scale datasets. Their study demonstrates that deep learning models outperform traditional machine learning methods, particularly in handling complex and high-dimensional data. The findings highlight the advantages of combining ML and DL approaches, supporting the idea of hybrid models to achieve improved accuracy and more robust phishing detection systems.

Wilk-Jakubowski et al.[11],(2025) present a comprehensive survey on phishing detection methods using machine learning and neural networks, covering research from 2017 to 2024. The study analyzes various techniques, datasets, and performance

trends, highlighting advancements and limitations in existing approaches. It identifies key research gaps, such as the need for more robust and hybrid models, thereby supporting further exploration and development in phishing detection systems.

Shilpa and Reddy.[12],(2025) propose a phishing detection approach that combines machine learning with a Deep Neural Network (DNN) using the PhishTank dataset. Their hybrid model leverages both traditional feature-based learning and deep feature extraction, resulting in improved accuracy and performance. The study demonstrates the effectiveness of integrating ML and DNN techniques, directly aligning with and supporting the proposed hybrid approach for phishing detection.

Sahingoz et al.[13], 2019 – Machine Learning Based Phishing Detection from URLs. This study used a dataset of ~70,000 URLs collected from PhishTank and legitimate sources. Feature extraction included lexical and host-based features. Models like Random Forest, SVM, and Naive Bayes were applied. Random Forest achieved the highest accuracy (~97%). This work relates to our project by highlighting the importance of feature engineering, which we extend using feature selection and deep learning models.

Le et al.[14], 2018 – URLNet: Learning a URL Representation with Deep Learning, Dataset includes large-scale URL data. CNN-based architecture extracts character-level features. Achieved high accuracy (~98%). This supports our approach of deep learning-based feature extraction.

Jain & Gupta[15], 2018 – Phishing Detection Using Visual Similarity.Used webpage screenshots dataset. Applied visual similarity techniques. Demonstrated improved detection of spoofed sites. This relates to feature diversity.

III. System Analysis

Phishing attacks are one of the most common cybersecurity threats, targeting users through malicious emails and websites. These attacks aim to steal sensitive information such as passwords, banking details, and personal data. Traditional detection methods struggle to keep up with evolving phishing techniques. There is a need for intelligent systems that can identify both known and unknown phishing patterns. The system must analyze features such as URLs, email content, and website structure. It should process large volumes of data efficiently. Machine learning and deep learning models can improve detection accuracy. Feature selection techniques are essential to identify the most relevant attributes. The system must ensure real-time detection and low false positives. Scalability is important for handling large datasets. Overall, the system requires a robust and adaptive approach for effective phishing detection.

Existing System

Existing phishing detection systems mainly rely on blacklists and rule-based approaches. These systems detect known phishing URLs by comparing them with stored databases. Some systems use basic machine learning models with limited features. However, these approaches are not effective against new or evolving phishing attacks. Feature extraction is often manual and limited. Existing systems have high false positive rates. They struggle with dynamic and obfuscated URLs. Real-time detection capabilities are limited. Many systems do not use deep learning techniques. Integration with advanced analytics is lacking. Overall, existing systems provide basic protection but lack adaptability and accuracy.

Disadvantages of Existing System

- Inability to detect zero-day phishing attacks
- High false positive and false negative rates
- Dependence on blacklists and predefined rules
- Limited feature selection techniques
- Lack of deep learning integration
- Poor scalability
- Inefficient handling of dynamic phishing strategies

Proposed System

The proposed system uses a hybrid approach combining machine learning and deep learning for phishing detection. It extracts features from URLs, emails, and website content. Feature selection techniques such as Chi-square or PCA are used to identify important attributes. Machine learning models like Random Forest and SVM are used for initial classification. Deep learning models such as CNN and LSTM are applied for advanced pattern recognition. The system can detect both known and unknown phishing attacks. It provides real-time detection with improved accuracy. The model is trained on large phishing datasets. It reduces false positives by selecting optimal features. The system adapts to evolving threats through continuous learning. Overall, it offers a robust and scalable phishing detection solution.

Advantages of Proposed System

- Detects both known and unknown phishing attacks
- Improved accuracy using ML and DL models
- Reduced false positives and negatives
- Efficient feature selection
- Real-time detection capability
- Scalable for large datasets
- Adaptive to evolving threats

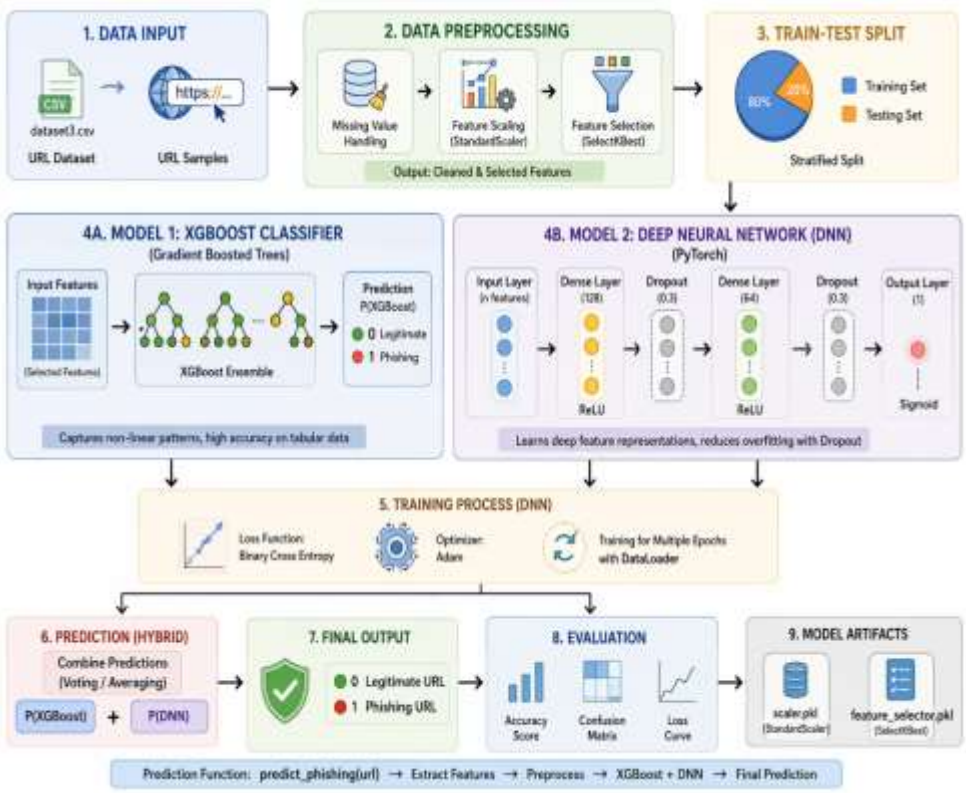
IV. Methodology

The methodology begins with collecting phishing and legitimate datasets from reliable sources. Data preprocessing is performed to clean and normalize the data. Feature extraction is applied to identify attributes such as URL length, domain age, and special characters. Feature selection techniques like PCA or Chi-square are used to select important features. The dataset is divided into training and testing sets. Machine learning models such as Random Forest and SVM are trained. Deep learning models like CNN and LSTM are also used. Model performance is evaluated using accuracy, precision, recall, and F1-score. Cross-validation is applied to improve reliability. Hyperparameter tuning is performed for optimization. The best model is selected for deployment. The system is tested for real-time phishing detection.

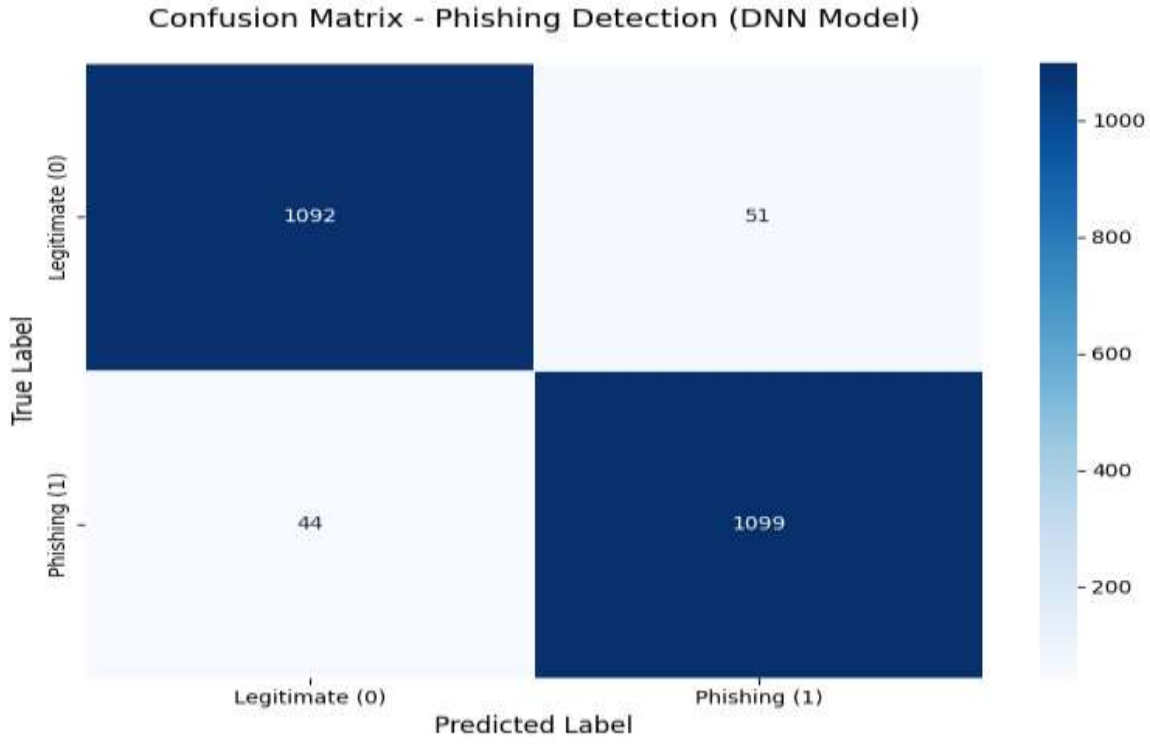
System Architecture

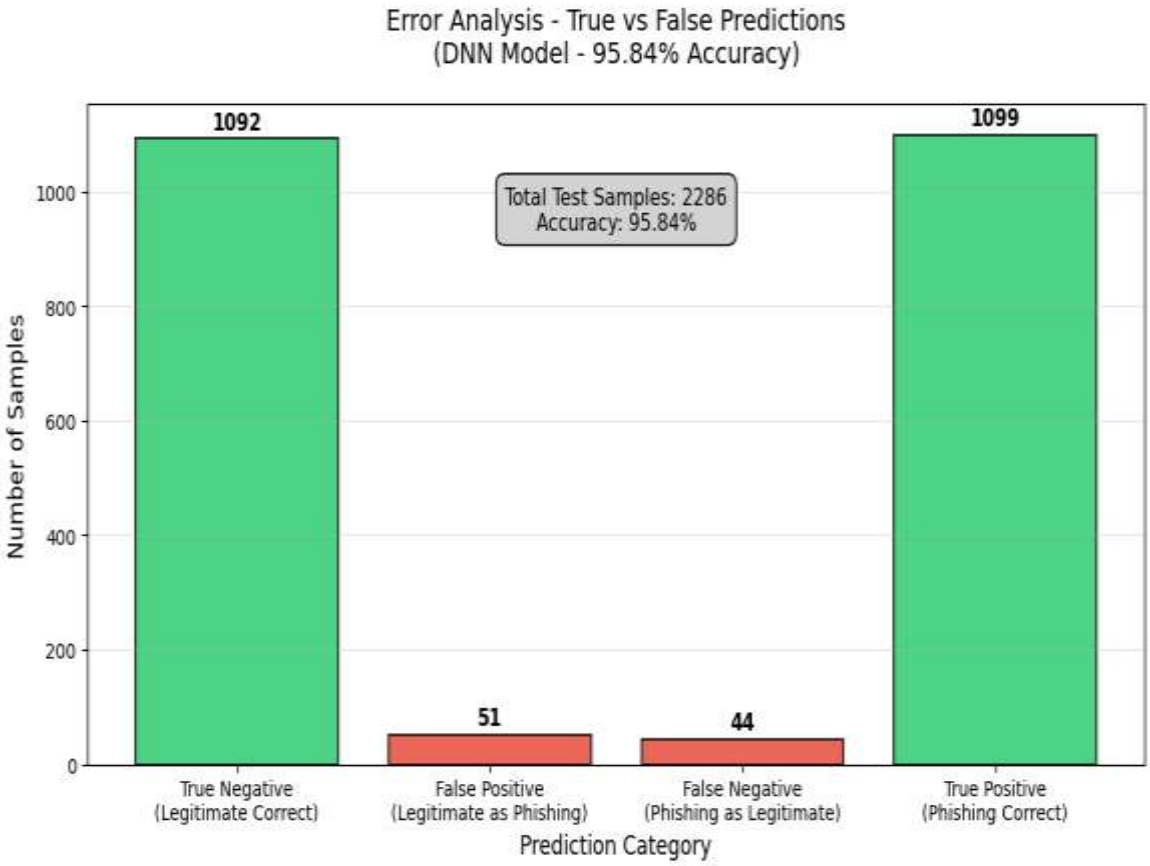
The system architecture consists of multiple layers. The data collection layer gathers URL, email, and website data. The preprocessing layer cleans and prepares the dataset.

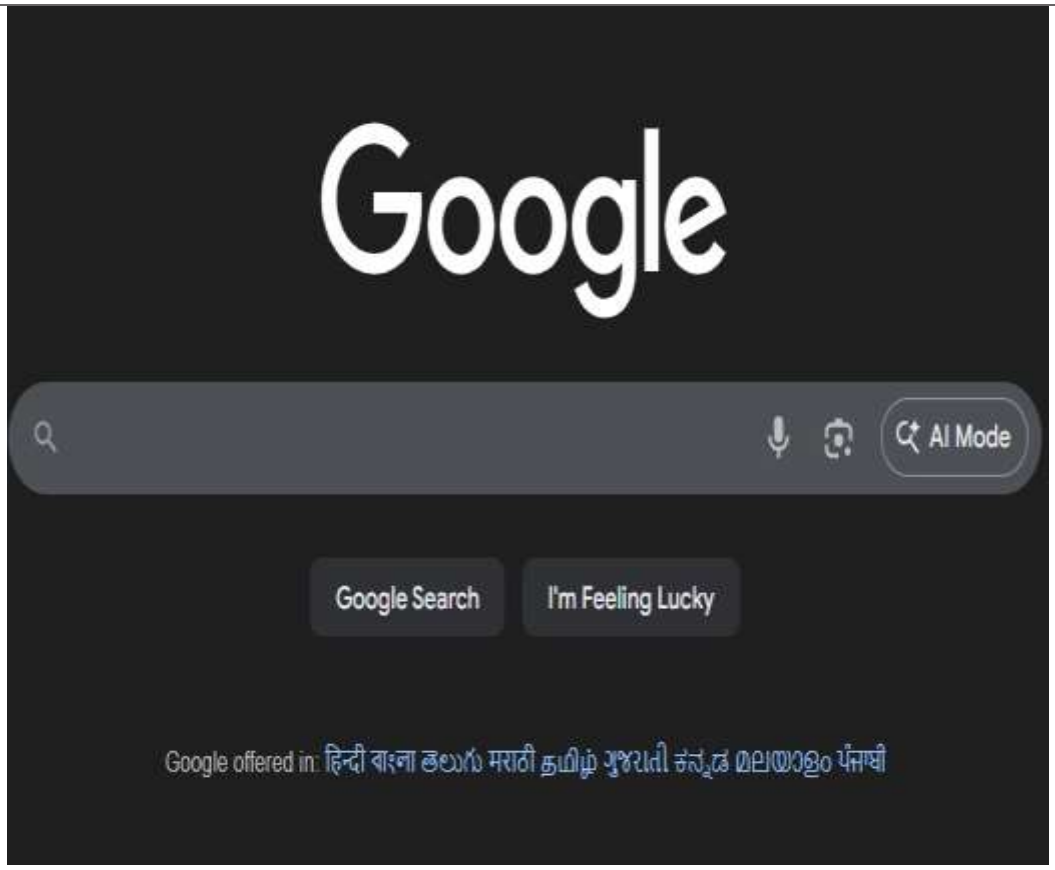
The feature extraction layer identifies relevant attributes. The feature selection layer selects important features. The model layer includes ML and DL algorithms for classification. The training module builds predictive models. The evaluation layer measures performance using metrics. The detection layer identifies phishing attempts. The database layer stores data and results. The user interface allows interaction with the system. The feedback layer updates the model with new data. Overall, the architecture ensures accurate and scalable phishing detection.

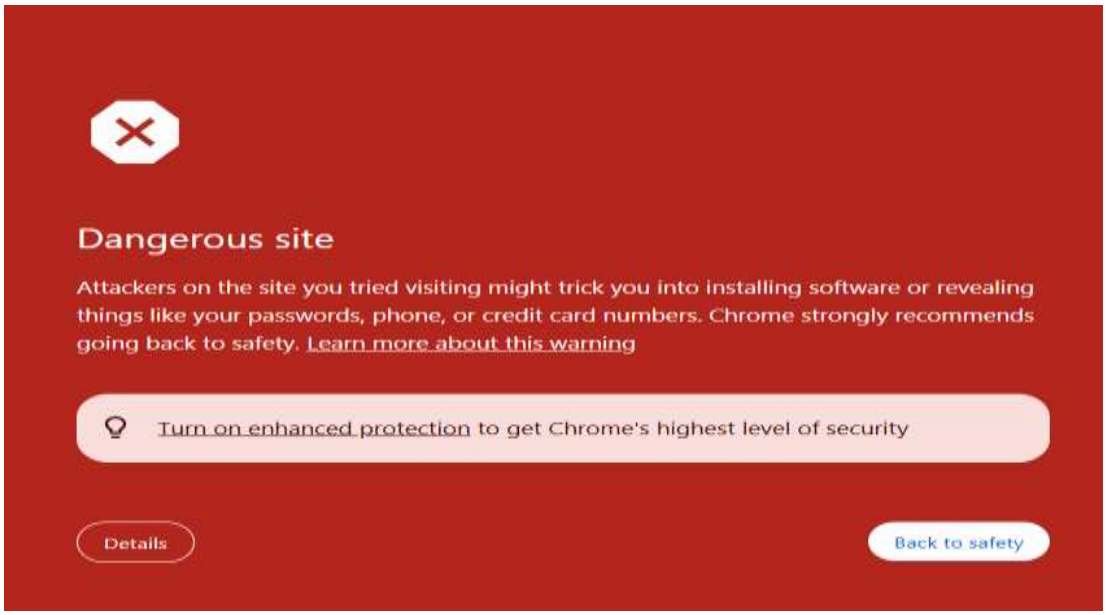


V. Result and Output


Confusion Matrix - Phishing Detection (DNN Model)
True Label
Predicted Label
Legitimate (0)
Phishing (1)
1092
51
44
1099
1000
800
600
400
200


Error Analysis - True vs False Predictions
(DNN Model - 95.84% Accuracy)
Number of Samples
Prediction Category
Total Test Samples: 2286
Accuracy: 95.84%
1092
51
44
1099
True Negative
(Legitimate Correct)
False Positive
(Legitimate as Phishing)
False Negative
(Phishing as Legitimate)
True Positive
(Phishing Correct)





VI. Conclusion

The proposed phishing URL detection system successfully demonstrates the effectiveness of a hybrid approach that combines machine learning and deep learning techniques. By integrating XGBoost with a Deep Neural Network (DNN), the model is able to capture both structured and complex patterns within URL-based features. The use of lexical and domain-specific attributes enables efficient and fast detection

without relying on webpage content, making the system practical for real-time applications. Overall, the project achieves reliable performance in distinguishing between legitimate and phishing URLs.

The implementation of a robust preprocessing pipeline, including feature scaling and feature selection, further enhances the model's accuracy and generalization capability. The DNN contributes to learning deeper representations of data, while XGBoost ensures strong performance on structured inputs. The ensemble strategy improves prediction consistency and reduces misclassification, making the system more dependable in handling diverse phishing scenarios.

Despite its strengths, the project also identifies certain limitations, such as dependence on URL-based features and lack of real-time threat intelligence integration. These challenges highlight opportunities for future improvements, including the incorporation of content-based analysis and adaptive learning mechanisms. Addressing these aspects can further strengthen the system's ability to detect advanced and evolving phishing attacks.

In conclusion, this project provides a scalable, efficient, and accurate solution for phishing detection using a hybrid modeling approach. It lays a solid foundation for further research and development in cybersecurity, particularly in building intelligent systems capable of combating modern phishing threats.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International

Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.