



## Phishguard Url Using Machine Learning

Deepthi P<sup>1</sup>, Aditya Vardhan G<sup>2</sup>, Likhitha K<sup>3</sup>, Nigama Sree Sai S.M<sup>4</sup>, Saleem Basha P<sup>5</sup>,  
Siva Bhargavi K<sup>6</sup>

<sup>1</sup>Faculty and <sup>23456</sup>Students

Dept. of Computer Science Engineering,  
Madanapalle Institute of Technology & Science  
Madanapalle, A P, India

<sup>1</sup>deepthisivaraj@gmail.com

<sup>2</sup>gadityavardhan143@gmail.com

<sup>3</sup>kamasanilikhitharoyal@gmail.com

<sup>4</sup>nigamasreesai@gmail.com

<sup>5</sup>psaleembasha6@gmail.com

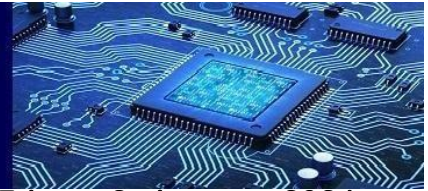
<sup>6</sup>sivabhargavisbr@gmail.com

**Abstract**—Phishing is still a major concern in the digital world because it uses fraud to obtain private information from people who are not careful. The advanced nature of phishing schemes is always improving, and traditional anti-phishing methods find it difficult to stay up. As a reaction, this study uses machine learning to provide an alternative way of identifying phishing websites. Specifically, we suggest a hybrid model that combines the Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), CatBoost, Gradient Boosting Classifier algorithms and we achieved good accuracy is 97.1%. Our methodology is centered on examining multiple aspects of URL importance to distinguish between genuine and fake websites. The hybrid approach efficiently detects fraudulent URLs by extracting and analyzing different data, such as domain age, URL structure, and linguistic characteristics. Several machine learning algorithms are used, which improves the flexibility and accuracy of the detection system. We show the effectiveness of our method in real-time phishing detection through thorough testing and assessment. The hybrid model achieves excellent accuracy and efficiency in identifying between genuine and fake websites, demonstrating higher efficiency in this regard. In today's digital environment, phishing attacks continue to pose a risk to users, but our all inclusive anti-phishing solution provides a promising response.

**Keywords**—Phishing, Fraudulent websites, Machine learning, Hybrid model, Multi-Layer Perceptron, Support Vector Machine, CiBoost, Gradient Boosting Classifier.

### 1. INTRODUCTION:

Attacks can be carried out by individuals like hackers, criminals, or white-capped attackers. To get to the computer is the goal [7]. Using the information within, or to gather personal data in other methods. The attacks began in 1988 and continue to this day as internet worms (known as the Morris Worm). The domains of fraud, forgery, threats, hacking, service blocking, malware applications, illegal digital materials, and social engineering are the principal targets of these attacks. Attackers exploit a variety of target users to obtain large amounts of money and/or



information. An estimated 500,000 individuals may fall victim to fraud, with approximately 300 million people in India being susceptible to phishing attempts. Shockingly, only a mere 7% of those affected report such crimes, often due to a lack of understanding of the repercussions.

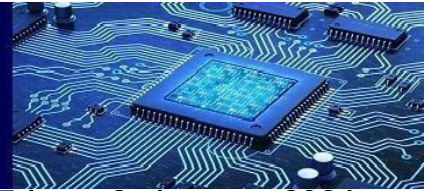
The internet offers a great deal of opportunity for criminals to carry out a variety of illegal activities. These activities include ransomware, worms, computer viruses, ransomware, online fraud, money laundering, theft, and extortion. [2] Hacking is a common threat that puts internet users at serious risk by enabling people to use computer information for malicious reasons. Concerns are heightened by the abundance of illegal material on the internet, especially when it comes to the influence it has on the next generation. Instead of depending just on websites that seem simple and secure, it is essential to find and identify those linked to these illegal activities in order to address these problems. It is imperative that users are aware of these malicious websites in order to protect themselves. Because they can propagate throughout computer networks, viruses risk not just the network but also private data kept on computers connected to it. To reduce these risks, using unapproved websites is strongly discouraged.

In order to protect computer systems from phishing, a common cyber threat, strong detection techniques are required. The past ten years have seen the development of numerous anti-phishing detection techniques as the field of cybersecurity has become a critical global concern. These approaches mostly focus on feature-selection techniques for machine learning applications, analyzing the structure of Uniform Resource Locators (URLs).

## **2. LITERATURE REVIEW**

To improve internet security,[1] suggest a hybrid model for phishing detection. By using bagging and boosting techniques, this approach leverages the strengths of multiple models. Data mining techniques are employed in the methodology to identify characteristics of phishing websites. When tested on a dataset consisting of 11,055 cases and 30 characteristics, the solution outperformed state-of-the-art methods with an accuracy of 99.25 percent. The study emphasizes how well the hybrid model works to mitigate threats to online security, especially when it comes to stopping phishing attacks that steal credentials.

[2] conducts research on countering the serious online threat of phishing attacks. In order to effectively protect against phishing URLs, the study uses a hybrid LSD model that combines logistic regression, decision trees, and support vector machines. This allows machine learning to be utilized. Evaluation metrics showing the effectiveness of the suggested method include specificity, recall, accuracy, precision, and F1-score. The hybrid model outperforms earlier models with accuracy rates of up to 95%, demonstrating its efficacy in tackling the critical problem of phishing in online security.



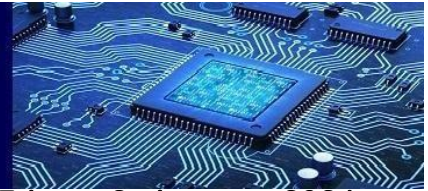
Phishing is an ever-changing online scam that requires advanced detection because traditional methods are insufficient.[3] introduced a novel approach to machine learning (ML) using Random Forest and Decision Tree classifiers, which proves to be an effective solution. This approach, which is described in the literature, uses a base dataset to train classifiers using feature extraction from web traffic data. Important elements include the domain, URL, and HTML/JS-based features. These classifiers are successful in differentiating between phishing and legitimate URLs, achieving high accuracies of 87.0% and 82.4%.

In this research,[4] address the significant risks to consumers' personal information and finances by presenting a novel technique for spotting phishing. SVM, Decision tree, Random Forest, and XGBoost algorithms are combined in the suggested framework to improve detection accuracy by combining elements from legitimate and phishing websites. By using different feature selection strategies and conducting a thorough examination, the hybrid methodology achieved an exceptional accuracy of 98.28%, surpassing the performance of state-of-the-art methods. This study adds to improved online security by offering a practical means of spotting and thwarting phishing attempts.

In the study referenced by [5], machine learning is used to detect phishing attempts, with a particular emphasis on using the SVM, Decision Tree and Random Forest algorithms. The work shows the efficacy of the Random Forest algorithm in improving detection techniques, attaining a low false positive rate and an accuracy of 97.14%. Through enhanced classifier performance, the research advances the detection of phishing websites by leveraging larger training datasets.

The thesis cited in [6] assembled a dataset of phishing websites from the UCI repository and used a variety of machine learning methods to analyze certain features, including AdaBoost, decision trees, random forests, and SVM. The most effective model for identifying phishing websites was found after a thorough evaluation. Two priority-based algorithms (PA1 and PA2) were subsequently created. With these algorithms and a new fusion classifier, the group was able to attain a remarkable 97% accuracy rate. This achievement exceeded earlier attempts to identify phishing websites.

A machine learning-based phishing detection system utilizing eight algorithms is proposed by[7]. The study, which focuses on the increase in cyberthreats targeting mobile devices, highlights the difficulties in identifying phishing attempts that imitate well-known websites. The results show exceptional performance, with faster detection times being shown by the Random Forest and Artificial Neural Network algorithms. When analyzing a single URL, the authors advise using the Random Forest algorithm for high accuracy and Artificial Neural Network for shorter training intervals. The study helps to improve the effectiveness of phishing detection systems in the face of growing cyberthreats.



A study on phishing detection with Extreme Learning Machine (ELM) is presented by [8]. The study tackles the difficulty of identifying phishing websites while highlighting the frequency and threat of phishing attacks. Thirty primary components are categorized by the suggested ELM approach using machine learning methods. Among the detection techniques are the evaluation of URL items, authority analysis of websites, and authenticity checks. By taking characteristics from URLs when users visit a website, the study seeks to identify phishing websites. In the end, users are alerted in advance to avoid unauthorized access to their personal data.

Using URLs from the University of California, Irvine Machine Learning Repository, a machine learning system was developed in the study cited as [9] to classify webpages. Four classifiers were used by the system: neural networks, decision trees, and SVM. Based on a trained dataset, the experimental results demonstrated the classifiers' ability to distinguish between authentic and fraudulent websites with an accuracy rate that surpassed 90%.

A Study [10] discusses the growing risk of online phishing attacks. These attacks, which are characterized as dishonest attempts to obtain private data, such as bank account information and passwords, present serious risks. The methods currently in use for detecting phishing URLs frequently demand significant time and financial commitments. In response, the study presents an effective method for identifying URLs that are phishing that is based only on URL features. The system effectively combats this ubiquitous online threat by detecting phishing websites with an impressive 96.35% accuracy rate using the SVM classifier.

According to research [11], the use of machine learning for phishing detection is enhanced by the Hybrid Ensemble Feature Selection (HEFS) framework. HEFS employs a two-stage process: first, a data perturbation ensemble is used to refine the primary feature subsets that are generated using the CDF-g algorithm. After that, baseline features are derived by a function perturbation ensemble. When combined with the Random Forest classifier, HEFS uses only 20.8% of the original features to distinguish between genuine and phishing websites, achieving an accuracy of 94.6%. The significant superiority of baseline features across classifiers demonstrates the efficacy of HEFS.

A machine learning-based method for phishing website detection in real-time is presented by Das Gupta, S., Shahriar, K.T., Alqahtani, H., Alsaman, D. and Sarker, I.H.[12] it focuses on URL and hyperlink-based features. New websites and zero-hour attacks are difficult for traditional anti-phishing techniques to handle. The paper addresses this by putting forth a hybrid feature-based approach that only uses client-side URL and hyperlink data to extract features. The XG Boost technique outperforms conventional methods with an experimentally demonstrated detection accuracy of 99.17%. The application of machine learning to counteracting emerging cyber threats is demonstrated by this research.



By utilizing potent machine learning algorithms, a cutting-edge method for identifying online phishing is presented in the study [13]. Using a huge collection of 20,000 website URLs, the study extracts 22 important features from each URL to address the growing sophistication of phishing attacks. To combat phishing websites that have text embedded in images, text-based analysis is also used. Phishing detection can be done accurately and successfully, according to experimental results. The maximum accuracy and precision rates that the XGBoost algorithm can achieve in training are 94%, and in testing, they are 91%. The study helps detect sophisticated phishing attacks early on by improving internet user security.

A. Ghimire, A. Kumar Jha, S. Thapa, S. Mishra and A. Mani Jha[14] study the widespread problem of phishing attempts on the internet and suggest using sophisticated algorithms to identify them. Their work investigates different approaches to machine learning-based phishing URL detection that combine network- and URL-based features. They also use techniques such as oversampling and undersampling to address dataset imbalance. The evaluation results show that machine learning effectively counters phishing attacks, with promising results in terms of precision, recall, F-score, and ROC AUC. By improving internet security and lessening the effects of cyberattacks, this research helps.

Provide a hybrid rule-oriented solution that incorporates six algorithm models to combat phishing attacks [15]. The study makes use of 37 features that were extracted from various techniques, including content analysis and blacklisting. When comparing accuracy, deep learning models outperform machine learning models; CNN achieves 97.945% accuracy, while MLP achieves 93.216%. Because of its high accuracy and efficiency, the study emphasizes the use of this solution in real-world settings to reduce phishing attempts.

### 3. PROPOSED APPROACH

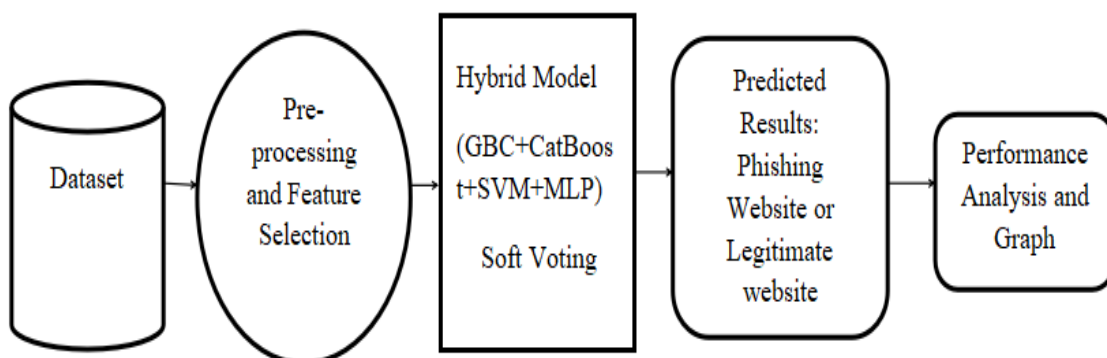




Fig.1. Architecture

### 3.1 Dataset

The initial phase of the system architecture for identifying phishing websites involves collecting a diverse dataset that comprises examples of both phishing and legitimate websites. There should be a range of phishing attack vectors included in this dataset, such as domain spoofing, email-based attacks, and deceptive website content. Web crawling tools, security firms, and publicly accessible repositories are some of the ways to obtain it. Class balance, dataset quantity, and data quality are carefully taken into account to ensure robust model training. We made use of a Kaggle dataset that contained 11,054 distinct data points, each of which has 32 features.

### 3.2 Pre-Processing and Feature Selection phase

The features include parameters such as

1. Index: Most likely, each entry in your dataset has a unique index.
2. UsingIP: Indicates if the website is using an IP address instead of a domain name.
3. LongURL: Indicates how long the URL is.
4. ShortURL: Denotes if the URL has been truncated.
5. Symbol@: The URL contains the '@' symbol.
6. Redirecting//: The URL's inclusion of '//' may suggest redirecting.
7. PrefixSuffix-: The URL contains prefixes or suffixes.
8. SubDomains: The total number of URL subdomains.
9. HTTPS: The utilization of the HTTPS protocol by the website.
10. DomainRegLen: The duration of registering a domain.
11. Favicon: The website has a favicon present.
12. NonStdPort: The communication using non-standard ports.
13. HTTPSDomainURL: The URL's domain portion contains HTTPS.
14. RequestURL: Indicates if the URL contains any requests.
15. AnchorURL: The URL contains anchor links.
16. LinksInScriptTags: How many links are contained within script tags.
17. ServerFormHandler: A server form handler is present.
18. InfoEmail: The existence of an email with information.
19. AbnormalURL: Indicates if the given URL is deemed abnormal.
20. WebsiteForwarding: The capability of sending webpages.
21. StatusBarCust: The status bar's customization.
22. DisableRightClick: shows whether the option to right-click is restricted.
23. Popup windows are used by usingPopupWindow.
24. IframeRedirection: Iframe redirection is present.
25. DNSRecording: DNS recording is present.
26. WebsiteTraffic: Data pertaining to traffic to websites.
27. PageRank: The website's page rank.



28. GoogleIndex: Listing in the index of Google.
29. LinksPointingToPage: The total count of outside links that lead to the page.
30. StatsReport: Indicates if a statistical report is accessible.
31. Class: The label here designates whether or not the website is considered to be phishing (1 for phishing, -1 for not phishing).

The label is the 'class' feature, where values represent whether or not a website is considered to be phishing (1 for phishing, -1 for not phishing). This dataset appears to offer a thorough overview of many different website aspects, which makes it useful for developing models for tasks related to phishing detection. Data cleaning strategies are utilized throughout the pre-processing and feature selection phase to address missing values, outliers, and inconsistencies. Understanding feature distributions and spotting possible patterns or anomalies are made easier with the help of exploratory data analysis. A vast range of features can be extracted from the dataset, including URL structural parameters like length and domain name as well as content-based characteristics like HTML analysis and keyword presence. Moreover, the best discriminative features for phishing detection are found using feature selection techniques like correlation analysis or recursive feature removal.

### 3.3 Hybrid Model

The hybrid model development, which uses the pre-processed dataset to train individual machine learning models like GradientBoostingClassifier, CatBoostClassifier, SVM, and MLP Classifier, is the system's central component. Techniques for hyperparameter tuning, like random or grid search, maximize each model's performance. Then, by utilizing a soft-voting method, these models are combined into a hybrid ensemble that enables the aggregation of class probabilities and sophisticated predictions. Strategies for managing class imbalance, such as oversampling and class weighting, are employed to ensure the model's performance in a range of scenarios. After the trained hybrid ensemble model is deployed, incoming website instances are categorized as either legitimate or phishing. Class probabilities are transformed into binary predictions using decision thresholds or probability cutoffs, and model interpretability techniques provide insight into the variables affecting predictions.

### 3.4 Performance analysis and Graph

Graph visualization and performance analysis are crucial for evaluating the model's effectiveness. Numerous metrics are computed to assess the model's performance, including ROC-AUC, F1-score, accuracy, precision, recall, and confusion matrix. Visual aids that show how the model behaves in various scenarios include ROC curves and precision-recall curves. Sensitivity analysis helps determine how robust the model is to changes in input features or data distributions, which guarantees its dependability in real-world scenarios.

## 4. METHODOLOGY



The study's hybrid model for phishing website detection combines the strengths of several machine learning classification algorithms, such as Multi-Layer Perceptron (MLP), Support Vector Machine (SVM), Gradient Boosting Classifier, and CatBoost Classifier. Through the integration of these disparate algorithms, the hybrid model leverages their unique strengths to augment efficiency and precision in the identification of phishing websites.

#### **4.1 Gradient boost classifier**

It is a technique for boosting. Gradient boosting is commonly applied to issues involving regression and classification. The primary goal of the boosting process is to create a strong model by uniting all of the weak learners. Gradient boosting is an extremely reliable method for creating prediction models. It is pertinent to multiple risk functions and optimizes the model's forecast accuracy.

#### **4.2 Cat Boost**

CatBoost is an open-source. Cat Boost can be used in ranking, recommendation systems, forecasting and even personal assistants. It is intended for use with very large numbers of independent features in regression and classification tasks. A gradient boosting variation that can handle both numerical and categorical information is called catboost.

#### **4.3 SVM**

One of the most widely used supervised learning techniques for both classifications and regression issues is Support Vector Machine. It's mostly applied to machine learning classifications challenges. SVM algorithm can be used for Face detection, image classification, text categorization, etc.

#### **4.4 Multi layer Perceptron**

Each layer of the MLP design is made up of networked nodes, or neurons. Inputs are received by each neuron, multiplied by matching weights, and then added together. Then, by adding non-linearity to the sum and applying an activation function, the neuron produces its output.

We take advantage of these different algorithms' complementary qualities by combining them into a hybrid model to produce a thorough anti-phishing solution. By examining various facets of URL significance, this method helps us identify fraudulent URLs with greater accuracy and efficiency. It also helps us spot phishing attempts.

### **5. EXPERIMENTAL RESULTS AND DISCUSSION**

We used a hybrid approach in our experimentation to enhance the effectiveness and performance of our machine-learning models. Using soft voting techniques, we combined the Gradient Boosting Classifier, CatBoost Classifier, SVM, and MLP Classifier into a single



ensemble model, which we refer to as (gbc+catboosting+svm+mlp). The ensemble model made use of the idea of "soft voting," which is a method of combining the predictions of several models by averaging their results. The ensemble model produces more reliable and accurate classifications when soft voting considers the degree of confidence in each model's predictions. By using this method to optimize the model's hyperparameters, we were able to find and pick the most informative features. The hybrid ensemble model demonstrated a remarkable accuracy rate of 97.1% in our experimental evaluations. This high degree of precision highlights how well our method works to identify phishing websites and shows how ensemble learning techniques can be used to improve the effectiveness of anti-phishing systems.

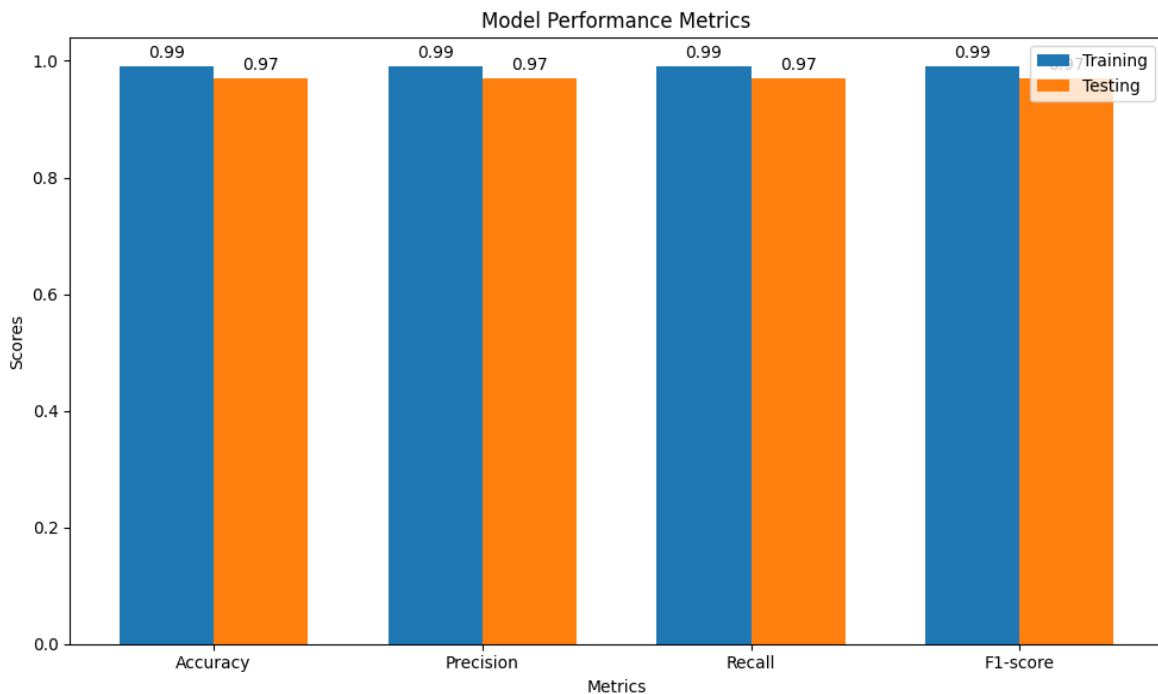
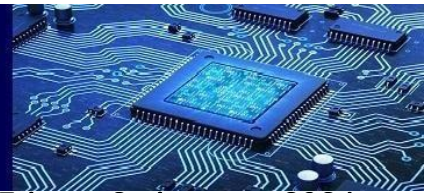


Fig.2.

These metrics are graphically represented by the bar chart that is shown in Fig.2, which provides a clear comparison between the training and testing datasets. Consistent performance between the training and testing phases is observed, indicating that the model is effective in differentiating between authentic and fraudulent websites, as indicated by high scores across all metrics. The corresponding metric scores are displayed at the top of each bar. The accuracy, precision, recall, and F1-score metrics are represented by the four bars.



	precision	recall	F1- score	support
-1	0.98	0.96	0.97	976
1	0.97	0.98	0.98	1235
accuracy			0.97	2211
Macro avg	0.97	0.97	0.97	2211
Weighted avg	0.97	0.97	0.97	2211

Table.1

Where,

- Precision=True Positives/(True Positives+False Positives).
- Recall=True Positives/(True Positives+False Negatives).
- F1-score=(2\*Precision\*Recall)/(Precision+Recall).
- Accuracy=(True Positives+True Negatives)/Total Instances.

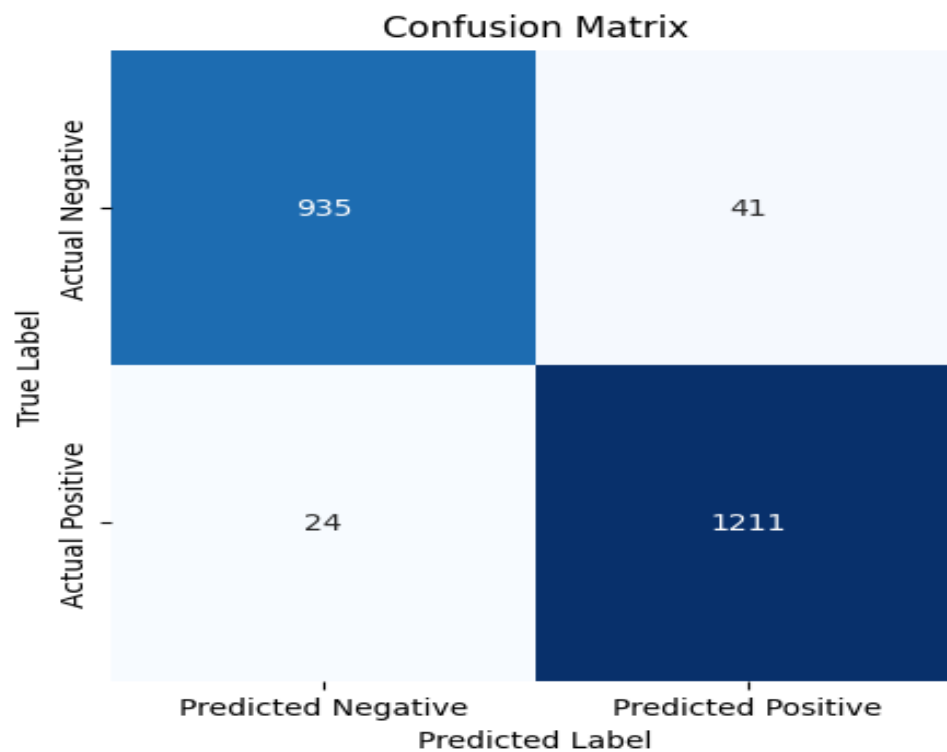




Fig.3.Confusion matrix

The hybrid model's confusion matrix for the testing dataset, which has 2211 records, represented in the figure3. It has 935 true positives, 1211 true negatives, 41 false positives, and 24 false negatives.

## 5. CONCLUSION

Phishing attacks need to be able to be predicted in real-time by an anti-phishing system. Expanding the reach of phishing site detection requires the prompt availability of a trustworthy anti-phishing solution. As of right now, our hybrid approach is limited to phishing website detection. We were able to combine The GradientBoostingClassifier, CatBoostClassifier, SVM, and MLP Classifier and achieve an amazing 97% detection accuracy with the lowest false positive rate by using soft voting techniques. While the use of URL lexical features has proven to be highly accurate, phishers have become adept at hiding URL destinations to prevent detection. Therefore, the best course of action is to combine these features with others, like host information. We intend to improve the phishing detection system in the future by turning it into a scalable web service. Online learning features will be incorporated into this service, allowing the system to adjust and pick up new phishing attack patterns instantly. Our goal is to effectively counter emerging phishing threats and improve the accuracy of our models by incorporating improved feature extraction techniques.

## 6. REFERENCES

- [1] Mirza, Amaad & Asghar, Sohail & Zafar, Ayesha & Gilani, Saira. (2016). A Hybrid Model to Detect Phishing-Sites Using Supervised Learning Algorithms. 1126-1133. 10.1109/CSCI.2016.0214.
- [2] A. Karim, M. Shahroz, K. Mustofa, S. B. Belhaouari and S. R. K. Joga, "Phishing Detection System Through Hybrid Machine Learning Based on URL," in IEEE Access, vol. 11, pp. 36805-36822, 2023, doi: 10.1109/ACCESS.2023.3252366
- [3] A. Mandadi, S. Boppana, V. Ravella and R. Kavitha, "Phishing Website Detection Using Machine Learning," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.
- [4] N. Tabassum, F. F. Neha, M. S. Hossain and H. S. Narman, "A Hybrid Machine Learning based Phishing Website Detection Technique through Dimensionality Reduction," 2021 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Bucharest, Romania, 2021, pp. 1-6, doi: 10.1109/BlackSeaCom52164.2021.9527806.
- [5] Mahajan, Rishikesh, and Irfan Siddavatam. "Phishing website detection using machine learning algorithms." International Journal of Computer Applications 181, no. 23 (2018): 45-47.



- [6] A. Lakshmanarao, P. S. P. Rao and M. M. B. Krishna, "Phishing website detection using novel machine learning fusion approach," 2021 ICAIS, Coimbatore, India, 2021, pp.1164-1169, doi: 10.1109/ICAIS50930.2021.9395810.
- [7] M. Korkmaz, O. K. Sahingoz and B. Diri, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis," 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2020, pp. 1-7, doi: 10.1109/ICCCNT49239.2020.9225561.
- [8] M. M. Vilas, K. P. Ghansham, S. P. Jaypralash and P. Shila, "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2019, pp. 384-389, doi: 10.1109/ICEECCOT46775.2019.9114695.
- [9] Kulkarni, Arun & L., Leonard. (2019). Phishing Websites Detection using Machine Learning. International Journal of Advanced Computer Science and Applications. 10. 10.14569/IJACSA.2019.0100702.
- [10] Banik, Bireswar, and Abhijit Sarma. "Phishing URL detection system based on URL features using SVM." International Journal of Electronics and Applied Research 5, no. 2 (2018): 40-55.
- [11] Chiew KL, Tan CL, Wong K, Yong KS, Tiong WK. A new hybrid ensemble feature selection framework for machine learning-based phishing detection system. Information Sciences. 2019 May 1;484:153-66.
- [12] Das Gupta, S., Shahriar, K.T., Alqahtani, H., Alsalman, D. and Sarker, I.H., 2022. Modeling hybrid feature-based phishing websites detection using machine learning techniques. Annals of Data Science, pp.1-26.
- [13] Shaukat, M.W.; Amin, R.; Muslam, M.M.A.; Alshehri, A.H.; Xie, J. A Hybrid Approach for Alluring Ads Phishing Attack Detection Using Machine Learning. Sensors 2023, 23, 8070.
- [14] A. Ghimire, A. Kumar Jha, S. Thapa, S. Mishra and A. Mani Jha, "Machine Learning Approach Based on Hybrid Features for Detection of Phishing URLs," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 954-959, doi: 10.1109/Confluence51648.2021.9377113.
- [15] Youness Mourtaji, Mohammed Bouhorma, Daniyal Alghazzawi, Ghadah Aldabbagh, Abdullah Alghamdi, "Hybrid Rule-Based Solution for Phishing URL Detection Using Convolutional Neural Network", Wireless Communications and Mobile Computing, vol. 2021, Article ID 8241104.