*Identification of Phishing Attacks using Machine Learning*

Nikhil Jindal   
 *Department of Computer Science & Engineering and Information Technology  
 Jaypee University of Information Technology*

Solan, India  
 <nikhiljindal1a@gmail.com>

Dhruv Rastogi  
*Department of Computer*

*Science & Engineering and Information Technology  
 Jaypee University of Information Technology*

Solan, India  
 [dr26dhruv@gmail.com](mailto:dr26dhruv@gmail.com)

Kartik Joshi   
*Department of Computer*

*Science & Engineering and Information Technology  
 Jaypee University of Information Technology*

Solan, India   
 <joshisaab2002@gmail.com>

Deepak Gupta

*Department of Computer*

*Science & Engineering and Information Technology*

*Jaypee University of Information Technology*

Solan, India

<deepak.vd@gmail.com>

*Abstract*— The Internet universal connectivity is both a boon and a breeding ground for phishing attacks, manipulating unsuspecting users into interacting with harmful links that redirect them to deceptive websites aiming to pilfer personal information. Phishing, a prevalent cyber threat, often uses cleverly disguised links in emails, texts, or social media that appear authentic, leading individuals to divulge sensitive details. Moreover, the unprecedented rise of artificial intelligence in recent years has also fueled a fresh wave of phishing attacks. These have become even more sophisticated, dangerous and prevalent. Therefore, there is a need to deploy advanced solutions that can automatically detect and prevent such phishing attacks. In this paper, the proposed solution analyzes URLs, considering factors like domain name, URL length, and the presence of suspicious keywords, seeking to differentiate between legitimate and phishing attempts. Further, various machine learning models are deployed to detect and classify phishing attacks to achieve an accuracy of 95.2%.

Keywords— Phishing Detection, Ensemble Learning, Feature Selection, Gradient Boosting, XGBoost.

# Introduction

Phishing, a prominent cyber threat, manipulates trust to steal sensitive data, making it challenging for individuals to distinguish legitimate website links. Utilizing various channels like emails, calls, and social media, these attacks lure victims through deceptive tactics, leading to severe repercussions like identity theft and financial loss. Proactive measures involve robust security protocols, staff education, and advanced cybersecurity tools. Proposing a web application specifically aimed at verifying website uniform resource location (URL) can empower users to discern between phishing and authentic sites, diminishing security risks for individuals and organizations, and ultimately addressing the growing threat of phishing attacks. Several reputable sources offer insights into phishing attacks: the Anti-Phishing Working Group (APWG) identifies hundreds of thousands to millions of phishing sites monthly. The FBI's IC3 [1] receives tens of thousands of annual phishing-related complaints. Cybersecurity Ventures estimates billions of daily phishing emails, totalling several billion attacks annually. Verizon's Data Breach Investigations Report [2] consistently underscores phishing as a major breach method.  
  
In the landscape of detecting URL-based phishing attacks, researchers have grappled with multifaceted challenges. Distinguishing authentic URLs from fraudulent ones presents a major hurdle, aggravated by the evolving sophistication of phishing tactics across various digital platforms. This complexity results in an increased susceptibility among users to fall victim to these attacks. However, countermeasures and solutions have been diligently crafted. Anti-phishing organizations like APWG and the efforts of cybersecurity firms have established tools and protocols aiming to identify and mitigate phishing risks. Awareness campaigns, advanced email filtering systems, and browser-based warning mechanisms have empowered users to discern potentially malicious URLs, substantially reducing the success rate of such phishing attempts. Constant research and innovative solutions continue to refine the defence against URL-based phishing assaults, creating a more resilient and informed digital ecosystem.

The objective of this paper is to encompass dataset collection, pre-processing, machine learning model selection, and web-based application development for phishing URL detection. The overarching motivation lies in safeguarding users against the pervasive threat of phishing attacks. These attacks, prevalent and menacing, jeopardize personal and corporate data by exploiting private information. Existing defence measures, predominantly user awareness and education, often prove insufficient, as even informed individuals can succumb to sophisticated phishing tactics. Moreover, these attacks are increasing at an unprecedented rate and therefore, there is a need to deploy big data analytics [3] to analyze the vast amount of information. Thus, the paper seeks to contribute to the field by providing advanced tools and methodologies capable of effectively identifying and thwarting these deceptive cyber assaults, offering enhanced protection to individuals and businesses in the digital realm.

# Related Work

This section describes current research developments on the use of machine learning algorithms for the classification of phishing websites.

Rao and Pais [4] proposed a technique based on how people behaved when they were on harmful websites. Heuristic filtering was added to automate the human behaviour of entering bogus credentials. They were able to obtain an accuracy of 96.38% by using this strategy.

The authors of [5] put forth a real-time system that incorporated natural language processing (NLP) characteristics. In order to differentiate between phishing and legitimate websites, they deployed seven different machine learning algorithms. Their results showed that RF with features based on NLP provided the best accuracy when it came to classifying web pages.

Xiaoqing et al. [6] offered a phishing webpage detection system that is automatic and clever. They employed Naive Bayes (NB) to classify the results after analysing the characteristics of the URL. Websites that appeared suspicious were processed and given a new classification using Support Vector Machine (SVM). They claimed that the system provided excellent detection accuracy in a short amount of time based on their findings.

A comparable strategy was applied in [7]. The authors described a procedure that combined a decision tree (DT) model with support vector machines (SVM). In order to produce the criteria for identifying phishing websites that target the banking domain, decision trees were utilized in conjunction with SVM for training purposes.

Joshi et al. [8] developed a system that used machine learning techniques and key aspects of URLs to identify and categorise phishing websites. The algorithms Relief and random forest (RF) outperformed other combinations, they discovered.

Wu et al. [9] provided a phishing detection tool that was created by fusing the webpage's URL with its source code. They employed the Levenshtein method to determine string similarity and SVM as a machine learning model in their suggested system for phishing webpage identification.

Tan et al. [10] presented a three-step strategy for identifying phishing websites called PhishWHO. The keywords were initially obtained from the websites using the N-gram technique. Using these keywords, a search engine was utilised to find the name of the target domain in the second stage. In the final phase, they employed a matching system for determining whether the webpage is legal.

Chiew et al. [11] image-based method for identifying fraudulent websites was suggested. They used the logo to search for it on Google Images in order to confirm the integrity of the website. They cross-referenced the domain name that Google returned with the webpage inquiry to identify phishing websites from authentic ones. Experiments were conducted to illustrate the usefulness of the recommended method.

Almseidin et al. [12] conducted a study in which they employed a variety of machine learning algorithms and features selection strategies to improve the efficacy of their system. A 48-character phishing dataset, comprising 5,000 trustworthy and 5,000 malicious URLs, was used for the testing. Their results showed that the RF method with only 20 features had the best accuracy.

A heuristic approach was proposed by Basnet and Doleck [13] using URL-based attributes. The trials employed a set of 138 features that were extracted from 16,000 phishing and 31,000 dangerous websites. These 138 features were divided into four categories: search engine, reputation, lexicon, and keywords. To achieve the categorization goal, they employed seven different machine learning techniques. Accuracy was higher with RF.

E. Gandotra and D. Gupta [14] looked into machine learning methods for phishing website detection. They brought in an extensive feature set that included HTML attributes, URLs, and web pages. Analysis of each individual feature was done before combining them for classification. The efficacy of URL-based features in webpage classification was highlighted by their findings. With RF as the best classifier, they were able to attain a remarkable 99.5% accuracy using their suggested method, with low false positive and false negative rates.

Yerima and Alzaylaee [15] presented a method for spotting phishing websites that is based on deep learning. Convolutional neural networks (CNNs) were utilised for this, and an evaluation dataset consisting of 4,898 phishing sites and 6,157 authentic websites was utilised. They discovered that their CNN-based method produced noticeably better outcomes than conventional machine learning.

In [16], the authors investigated the role of feature selection strategies in improving the efficacy and efficiency of phishing webpage detection. They used a dataset comprising 6,157 benign and 4,898 phishing webpages with 30 features to compare different machine learning algorithms. With RF emerging as the best-performing algorithm both before and after feature selection, their research demonstrated the beneficial effects of feature selection. This method not only increased accuracy but also sped up the process of building the model, highlighting the importance of feature selection in early phishing detection

In [17], the authors investigated a number of machine learning and ensemble techniques during their thorough investigation of the detection of phishing webpages. A dataset comprising 30 features was used in their research, which included 6,157 benign and 4,898 phishing webpages. The most accurate method was found to be stacking, which achieved 96.987% accuracy in phishing webpage detection.

The aforementioned studies suggest that researchers and practitioners are developing solutions to detect phishing attacks early to mitigate the risks involved with such attacks. This paper aims to develop a behavioural analysis-based approach that considers user behaviour patterns and anomalies to identify potential phishing URLs. By analysing user interactions and behaviour, such as mouse movements and keystrokes, detection systems can identify suspicious URLs that exhibit anomalous behaviour.

# Methodology

The methods used to distinguish between benign and malicious webpages is described in this section. The suggested architecture of the phishing detection system is shown in Fig. 1. It commences with dataset acquisition and feature extraction, leading to system training and evaluation. If the assessment meets criteria, the system is implemented; otherwise, it undergoes further training. Utilizing classifiers like random forest, support vector machine, decision tree, logistic regression, gradient boosting, extreme gradient boosting, multilayer perceptron, and naive bayes, it categorizes URLs as legitimate or potential phishing links based on input. In a study evaluating 7900 malicious and 5800 legitimate sites, this method outperformed recent approaches in identifying phishing URLs, signifying the prowess of machine learning in detecting deceptive links.

Feature extraction is pivotal in reducing the dimensionality of considered features, aiming to derive a subset of vital or principal features. Algorithms for feature extraction endeavour to condense a high-dimensional feature set into a more manageable and less redundant vector, employing methodologies like principal components analysis (PCA). Fig. 2 represents the top 12 features after feature selection using PCA that ranks different features within a machine learning model. HTTPS emerges as the most crucial feature, while PageRank is deemed the least important. Feature importance is determined based on the impact of a feature's removal on the model's overall performance, providing valuable insights into the significance of individual features within the model.

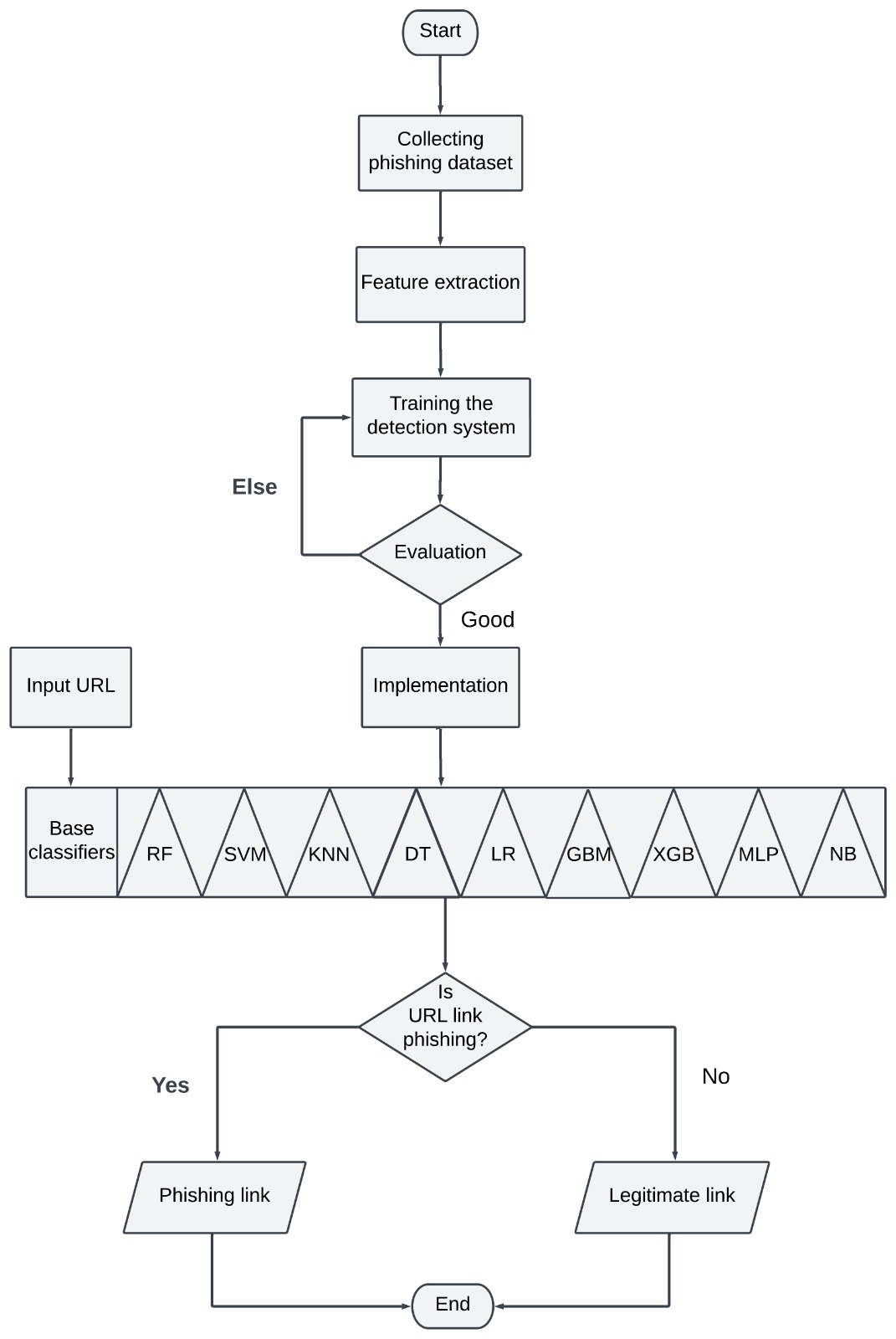


Fig. 1.Workflow of methodology

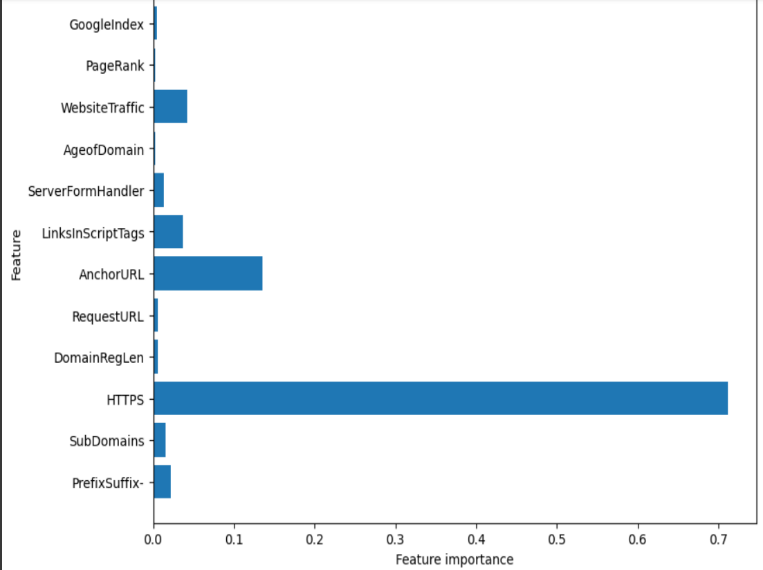


Fig. 2.Top 12 features after feature selection.

# Evaluation and Validation

An approach known as ten-fold cross-validation has been employed in the experimentation. The way this method operates is by randomly splitting the original dataset. First, ten equal portions of the original dataset are divided. Nine of the pieces are then utilized for training, while one is used for testing. Ten times, the identical procedure is carried out using various combinations. Algorithm performance is evaluated using the averaged outcome. We have utilized a range of evaluation factors to assess machine learning algorithms. These include accuracy, precision, F-measure, accuracy, false positive rate (FPR), and Matthews correlation coefficient (MCC)

• **TPR:** It is also referred to as recall and is the percentage of phishing URLs that are correctly recognized.

*TPR* **=**  (1)

**• FPR:** It's the percentage of benign webpages that are misidentified.

*FPR* **=**  (2)

**• Precision:** It is a degree of exactness.

*Precision* **=**  (3)

**• F-Measure:** Its definition is the precision and recall harmonic mean.

*F - Measure* **=**  (4)

**• Accuracy (%):** It is the proportion of safe webpages and phishing sites that are successfully identified.

*Accuracy* **=**  (5)

**• MCC:** It is used in assessing how well machine learning algorithms perform in binary classification. It takes values between -1 and + 1 and calculates the correlation between the actual and anticipated labels.

*MCC* *=*

# Results

In this section, the results of the experiments are compared, and visually represented for better understanding. All nine classifiers are utilized to categorize phishing websites, evaluating a dataset encompassing 12 distinct features. Table I displays the weighted average outcomes for accuracy(5), recall, MCC(6), and F1-measure(4) for each of the nine machine learning methods that use 12 features. Table IandFig. 3 show that gradient boosting classifier delivers the best accuracy of 95.2% followed by XGBoost classifier, RF and multi-layer perceptron with 95%, 94.9% and 94.7% accuracy respectively.

TABLE I. Classification results of different ML models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy**  **(%)** | **F-Measure** | **Precision** | **Recall** | **MCC** |
| Gradient Boosting Classifier | 95.2 | 0.951 | 0.967 | 0.965 | 0.902 |
| XGBoost Classifier | 95.0 | 0.949 | 0.976 | 0.957 | 0.970 |
| Random Forest | 94.9 | 0.948 | 0.969 | 0.968 | 0.979 |
| Multi-layer Perceptron | 94.7 | 0.946 | 0.958 | 0.959 | 0.959 |
| Decision Tree | 94.4 | 0.943 | 0.970 | 0.969 | 0.979 |
| SVM | 93.9 | 0.938 | 0.946 | 0.943 | 0.907 |
| KNN | 93.8 | 0.937 | 0.961 | 0.961 | 0.933 |
| Logistic Regression | 91.8 | 0.916 | 0.923 | 0.920 | 0.854 |
| Naive Bayes Classifier | 59.1 | 0.555 | 0.762 | 0.636 | 0.389 |

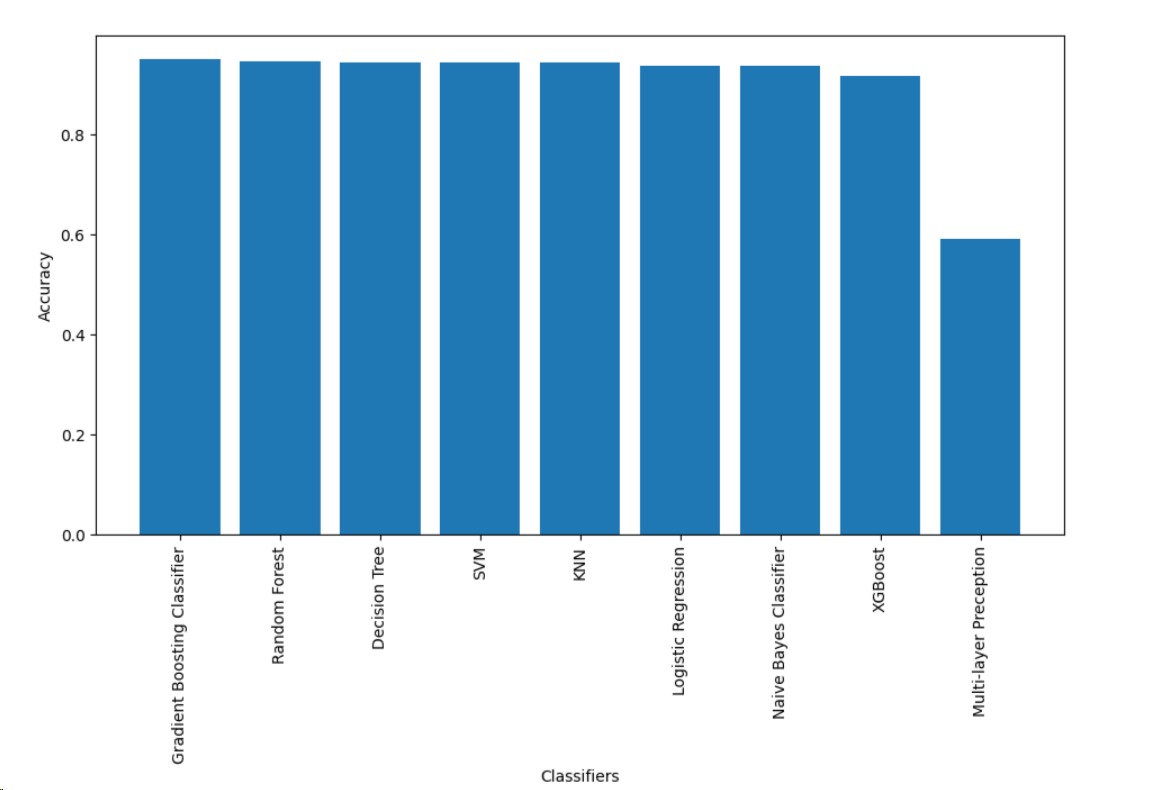


Fig. 3. Comparison of classification models on the basis of accuracy.

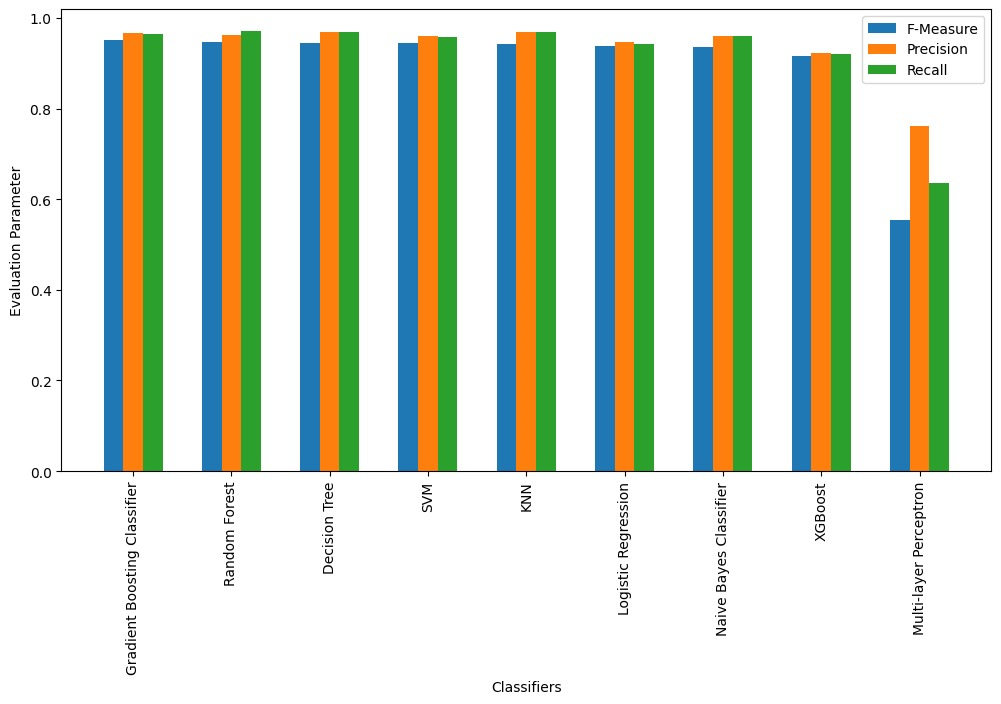


Fig. 4. Comparison of classification models based on F-Measure, precision & recall.

Fig. 5 illustrates the comparative performance of diverse machine learning algorithms on a specific dataset, evaluated using the MCC. The analyzed algorithms, encompassing logistic regression, KNN, NB, DT, RF, gradient boosting, and XGBoost, exhibit varying performance levels on both training and test sets. Notably, the gradient boosting algorithm demonstrates the highest MCC, while the DT algorithm records the lowest MCC. The MCC, widely acknowledged in machine learning, provides a balanced evaluation for binary and multiclass classifications by considering true and false positives and negatives. Ranging from -1 to +1, an MCC value of +1 signifies impeccable predictions, 0 indicates an average random prediction, and -1 represents an inverse prediction, making it a valuable metric for assessing classification quality.

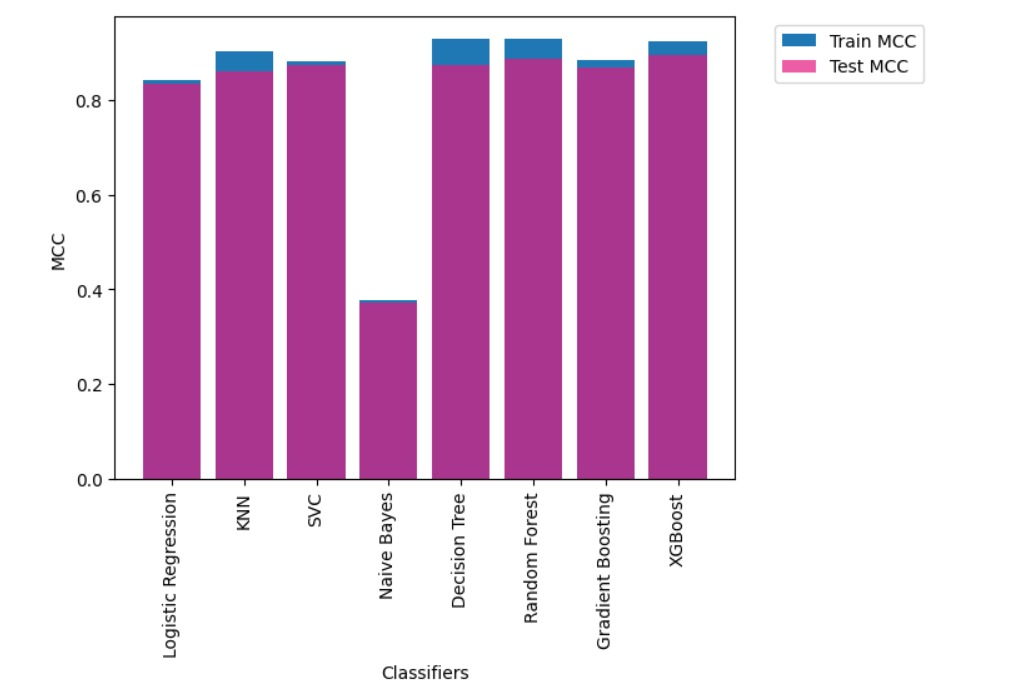


Fig. 5. Comparison of classification models on the basis of MCC.

Fig. 6 illustrates the receiver operating characteristic (ROC) curve, commonly employed to evaluate binary classifier models. It showcases the TPR(1) against the FPR(2) at different threshold levels, offering insights into the model's performance of the gradient-boosting classifier. The area under the curve (AUC), measuring 0.97 in this scenario, signifies the classifier's overall efficacy. A higher AUC denotes better model performance, and in this case, an AUC of 0.97 indicates a commendable performance level for the classifier. This analysis aids in comprehending the trade-offs between sensitivity and specificity in the model's predictions, offering a crucial evaluation metric for classifier performance.

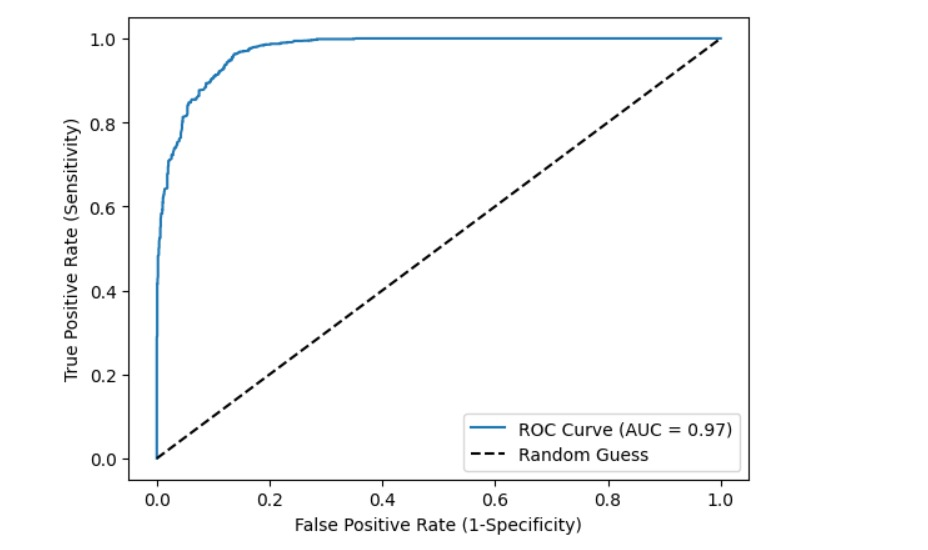


Fig. 6. ROC curve for gradient boosting classifier.

# CONCLUSION

The landscape of detecting and preventing URL-based phishing attacks is rife with challenges and opportunities. Phishing, a prevalent cyber threat, employs deception to manipulate individuals into interacting with harmful links, leading to potential data breaches and severe repercussions for both individuals and organizations. The utilization of machine learning algorithms in phishing detection offers a promising solution, showcasing its efficacy in distinguishing between legitimate and malicious URLs.

The development of machine learning models using various classifiers such as RF, SVM, DT, and gradient boosting has demonstrated differing degrees of success in identifying phishing links. Further, evaluative metrics like MCC and ROC curves provide insights into the performance of these models, with the MCC showcasing the overall effectiveness of the classification and the ROC curve assessing the trade-offs between true positive and false positive rates. Moreover, feature extraction techniques, such as assessing the importance of features like HTTPS and PageRank contribute significantly to understanding the key discriminative elements in identifying phishing URLs. Additionally, gradient boosting classifier provides the highest accuracy of 95.2% followed by XGBoost classifier, random forest and multi-layer perceptron with 95%, 94.9% and 94.7% accuracy respectively.

This multifaceted approach emphasizes the critical need for sophisticated tools and methodologies to combat phishing attacks effectively. The comprehensive evaluation and understanding of various models, metrics, and feature importance enable the development of robust systems for distinguishing between authentic and fraudulent URLs. The dynamic nature of cyber threats necessitates a continuous enhancement of these systems to adapt to evolving phishing techniques. In the future, the proposed solution is anticipated to be established as a cloud service for detecting phishing URLs.

##### References

1. FBI Internet Crime Complaint Center (IC3), "Internet Crime Complaint Center (IC3)," Accessed: October 31, 2023. Available: https://www.ic3.gov/
2. Verizon, "Verizon's Data Breach Investigations Report," Accessed: October 31, 2023. Available: https://enterprise.verizon.com/resources/reports/dbir/
3. D. Gupta and R. Rani, “A study of big data evolution and research challenges”, Journal of Information Science, pp.322–340, 2019.
4. S. R. Rao and A. R. Pais, "Detecting phishing websites using automation of human   behavior," *in the 3rd ACM Workshop on Cyber-Physical System Security*, ACM, pp. 33-42, 2017.
5. O. K. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Systems with Applications*, vol. 118, pp. 345-357, 2019.
6. X. Gu, H. Wang, and T. Ni, "An efficient approach to detecting phishing web," *Journal of Computer and Information Systems*, vol. 53, pp. 5553-5560, 2013.
7. M. Moghimi and A. Y. Varjani, "New rule-based phishing detection method," *Expert Systems with Applications*, vol. 54, pp. 231-242, 2016.
8. A. Joshi, P. Pattanshetti, and R. Tanuja, "Phishing Attack Detection using Feature Selection Techniques," *in the International Conference on Communication and Information Processing (ICCIP)*, Nutan College of Engineering & Research, 2019.
9. C. Y. Wu, C. C. Kuo, and C. S. Yang, "A phishing detection system based on machine learning," *in the International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, pp. 28-32, 2019.
10. C. L. Tan, K. L. Chiew, and K. Wong, "PhishWHO: phishing webpage detection via identity keywords extraction and target domain name finder," *Decision Support Systems*, vol. 86, pp. 18-27, 2016.
11. K. L. Chiew, E. H. Chang, and W. K. Tiong, "Utilization of website logo for phishing detection," *Computers & Security*, vol. 53, pp. 16-26, 2015.
12. M. Almseidin, A. Zuraiq, M. Al-kasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 13, no. 6, pp. 171-183, 2019.
13. R. B. Basnet and T. Doleck, "Towards developing a tool to detect phishing URLs: a machine learning approach," *in the IEEE International Conference on Computational Intelligence & Communication Technology*, pp. 220-223, 2015.
14. E. Gandotra and D. Gupta, "Improving spoofed website detection using machine learning", *Cybern. Syst.*, vol. 52, no. 2, pp. 169-190, Oct. 2020
15. S. Y. Yerima and M. K. Alzaylaee, "High accuracy phishing detection based on convolutional neural networks," *arXiv preprint arXiv*:2004.03960, 2020.
16. E. Gandotra and D. Gupta, “An Efficient Approach for Phishing Detection using Machine Learning”, *Algorithms for Intelligent Systems*, Springer, Singapore, 2021
17. D. Gupta, E. Gandotra, Y. Mohan and S. Singh, “Analysis of Ensemble Methods for Phishing Detection”,  *Intelligent Multimedia Signal Processing for Smart Ecosystems*, Springer, Springer, Cham. https://doi.org/10.1007/978-3-031-34873-0\_4.