IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

www.ijitce.com

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 13, Issue 2, 2025

Phishing Detection System Through Hybrid Machine Learning Based on URLs

P. Siva Srinivasarao[1], Gontla Lakshmi Veera Lahari[2], Marella manoghna[3], Jetti Veeranjaneyulu[4], Muttineni Trinadh[5]

[1]Assistant Professor, Department of Computer Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

[2,3,4,5] Students, Department of Computer Science Engineering, Chalapathi Institute of Engineering and Technology, Chalapathi Rd, Nagar, Lam, Guntur, Andhra Pradesh- 522034

Email id: thrishiva123@gmail.com[1], gontlalvlahari2003@gmail.com[2], Manoghna27@gmail.com[3], veeranjaneyulu81584@gmail.com[4], trinadh.m029@gmail.com[5]

Abstract:

Phishing attacks represent one of the most dangerous forms of cybercrime, exploiting email and websites to deceive individuals and obtain sensitive information. With phishing incidents on the rise, there is a significant need for robust defense mechanisms that leverage advanced machine learning techniques. This study proposes a hybrid LSD (Logistic Regression, Support Vector Machine, and Decision Tree) model aimed at improving phishing detection accuracy and efficiency. The LSD model utilizes both soft and hard voting mechanisms to combine the strengths of the constituent algorithms. For feature selection, we applied the canopy feature selection method, followed by cross-validation and hyperparameter tuning via Grid Search to optimize model performance. The effectiveness of the proposed approach is measured using evaluation metrics such as precision, accuracy, recall, F1-score, and specificity. Comparative analyses demonstrate that the hybrid LSD model significantly outperforms standalone classifiers, including Multinomial Naive Bayes, highlighting its potential in proactive phishing defense.

Keywords: Phishing Detection, Machine Learning, Hybrid Model, Logistic Regression, Support Vector Machine, Decision Tree.

1.Introduction

The internet provides many advantages in different fields of life. In the field of information search, the Internet has become a perfect opportunity to search for data for educational and research purposes. Email is a messaging source in fast way on the Internet through which we can send files, videos, pictures, and any applications, or write a letter to another person around the world. E-commerce is also used on the internet. People can conduct business and financial deals with customers worldwide through he-commerce. Online results are helpful in displaying results online and have become a more useful source of the covid-19 pandemic in 2020. Many classes and business meetings are performed online, which requires time and is fulfilled through the internet. Owing to the increase in data sharing, the chances of loss and cyber-attack also increase. Online shopping is the biggest Internet use that helps traders sell projects online worldwide. Amazon operates a large online sales system. Fast communication is performed through the Internet, which is currently used through Facebook, Instagram, WhatsApp, and other social networks, making communication fast and easily available. Therefore, it is necessary to maintain a privacy policy in which communication and its user scan not bed effective. The Internet provides a great opportunity for attackers to engage in criminal activities such as online fraud, malicious software, computer viruses, ransomware, worms, intellectual property rights, denial of service attacks, money laundering, vandalism, electronic terrorism, and extortion. Hacking is a major destroyer of the Internet through which any person can hack computer information and use it in different ways to harm others. Immorality, which harms moral values, is a major issue for the younger generations. Detecting these websites rather than websites that appear simple and secure, will help people. Therefore, an awareness of these websites is necessary. Viruses can damage an entire computer network and confidential information by spreading to multiple computers. It is not suitable to use unauthorized websites on the internet. Phishing detection is required for all of these aspects to secure our computer system.

2. Literature Review

Phishers are those attackers who plan phishing attacks. They create phishing websites that look similar to the legitimate ones to emulate original websites for stealing user's personal and sensitive details. The information achieved by attackers are often utilized to access users confidential accounts such as twitter, face book, email, bank etc. Many users put up with identity theft and financial losses due to the increasing number of phishing attacks [1]. Due to the advancement in technology, security concerns have been increasing for various sectors like banking, edu-cation, entertainment and so on. According to Gartner, U.S. banks and credit card companies have lost 2.8 billion dollar annually due to the theft through phishing attacks [2].According to APWG report [3], 165772 phishing sites have been detected in the first quarter of 2020 and 162155 phishing sites have been identified in last quarter of 2019 (see Fig.1).It is a matter of great concern that attackers focus on acquiring access to corporate accounts that pertain sensitiveand confidential financial information .There have been few works on phishing website detection. Some of the works are based on Blacklist and White list-based technique [4]. Some are based on Content-based approach [5]. Some are based on Visual similarity-based techniques [6]. Some are Heuristics and machine learning-based techniques [7]. Abdelhamid et al. [2] examined the problem of website phishing attack using Multi-label Classifier based Associative Classification (MCAC). Nearly 94.5% accuracy was obtained using MCAC. However, their model used a data set containing only 601 legitimate and 752 phishing websites. Only 16 features were utilized to detect phishing attack whereas there are other important features that could havebeen used for precise detection. In [8], the authors applied only Naive Bayes and sequential minimal optimization on two feature subsets (CFS and consistency subset) and could.

3. Material And Methods

Phishing detection based on URLs proposed in this study. The classification of phishing URLs was implemented using machine learning algorithms. Cybercrimes are growing with the growth of Internet architecture worldwide, which needs to provide a security mechanism to prevent an attacker from getting confidential content by breaching the network through fake and malicious URLs. Aphishing data set was used toper form the experiments. The dataset is in the form of data vectors that require null-value removal to remove unnecessary empty values. Multiple machine learning algorithms, such as decision tree (DT), linear regression (LR), naive Bayes (NB), random forest (RF), gradient boosting machine (GBM), support vector classifier

(SVC), K-neighbors classifier, and the proposed hybrid model (LR+SVC+DT) LSD with soft and hard voting were used based on functional features, as shown in Figure.

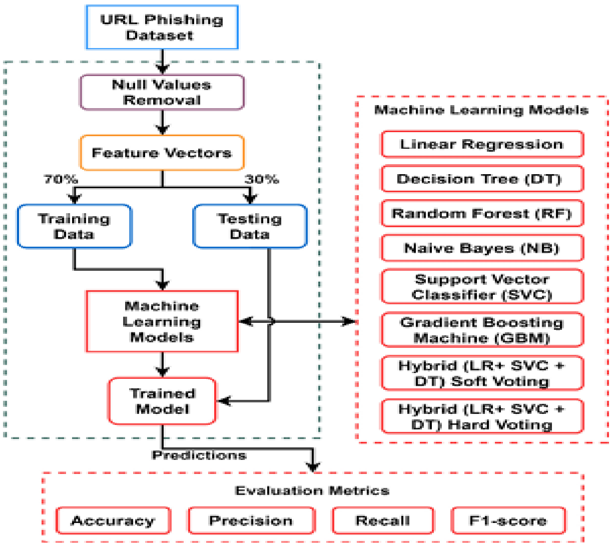


Figure: Classification of phishing URLs proposed based on designed methodological structure

DECISION TREE

The decision tree classifier (DTC) is a non-parametric method used for classification and regression. The decision tree classifier recursively partitions the given dataset of rows by applying the depth-first greedy method or the breadth-first approach until all data parts relate to an appropriate class. A decision tree classifier structure was created for the root, internal, and leaf nodes. Tree construction was used to classify unknown data. At each inner node of the tree, the best separation decision is made using impurity measures the leaves of the tree were created from the class labels in which the data objects were gathered. The DT (decision tree) classification procedure is implemented in two stages: tree building and tree pruning It is very tasking and computationally fast because the training dataset is frequently traversed. For a single attribute, entropy is mathematically expressed as

$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(1)}$$

The Entropy can be numerically expressed for various characteristics as follows:

$$E(T, X) = \sum_{c \epsilon X} p(c) E(c) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2)}$$

IG is defined mathematically by:

$$IG(T, X) = E(T) - E(T, X) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(3)}$$

support vector machine (SVM):

A support vector machine (SVM) is a supervised machine learning algorithm defined by a separating hyperplane between different classes. In other words, given the labeled training data (supervised learning), the algorithm outputs an optimal hyperplane that categorizes new test data based on the training data. Support vector machine (SVM) can be used for both classification and regression. However, it is mostly used in classification problems, where it provides the best accuracy between two classes.

4. EXPERIMENTAL RESULTS

The decision tree algorithm depends on tree-based architecture, which consists of several internal nodes and leaves that carry data according to the patterns found in the dataset. The sk learn library was used to access the tools for implementing the decision tree algorithm. Table presents the results of the proposed decision tree algorithm with the phishing data set to classify URLs in binary classes of 0 and 1. Decision tree.

Table: Results for the performance of the decision tree model.

max depth	Accuracy	Precision	Recall	Specificity	F1-score
0	94.9	95.46	95.41	94.25	95.44
5	92.07	89.67	96.97	85.85	93.18
10	94.3	94.59	95.23	93.09	94.92
20	95.38	95.7	96.06	94.53	95.88
30	95.41	95.8	96	94.66	95.91

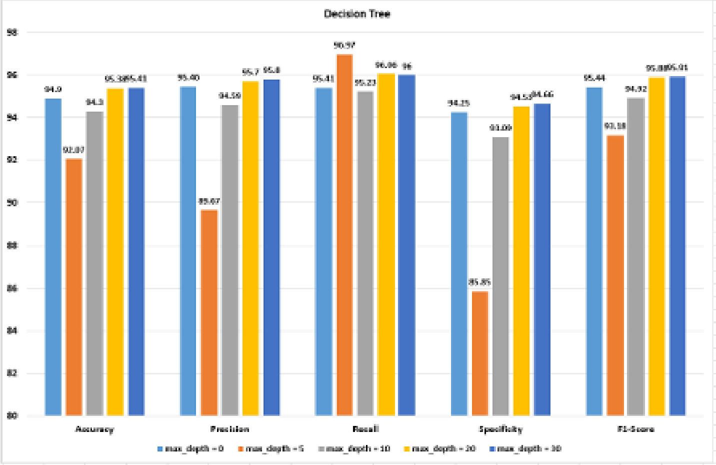


Figure: Experimental results of the decision tree model.

algorithms consist of many parameters, but the most effective parameter that affects the training and prediction accuracy of the model is max_depth. This parameter defines the depth of the tree in terms of its level.

Table: Results for the performance of the naive bayes model.

Accuracy	Precision	Recall	Specificity	F1-score
88.39	94.92	83.71	94.32	88.96

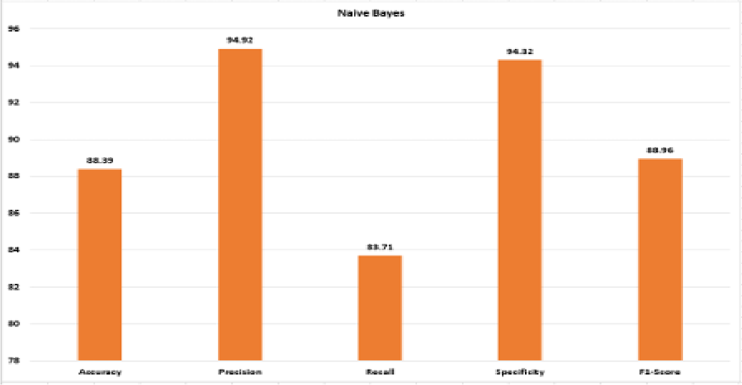
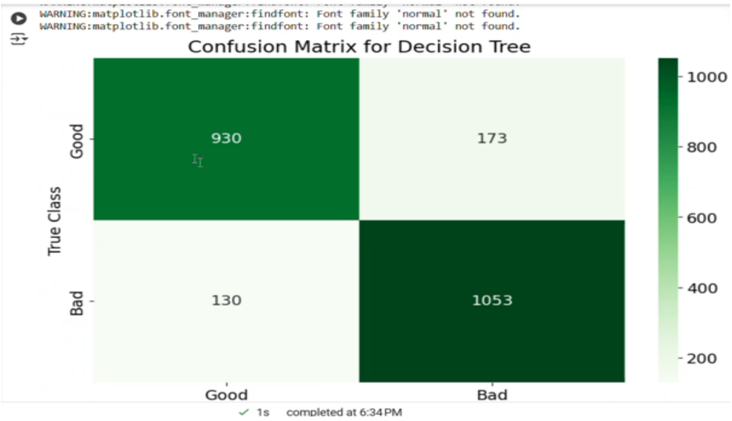


Figure: Experimental results of the naive bayes model.



```
WARNING:matplotlib.font_manager:findfont: Font family 'normal' not found.
WARNING:matplotlib.font_manager:findfont: Font family 'normal' not found.
```

Conclusion:

Instances of internet fraud are increasing nowadays and occur via different methods, among which phishing is one of the most popular. Scammers use fake websites to steal data, with URLs resembling those of original and legitimate websites. To address this issue, in the present study, a model was developed to detect phishing and legitimate websites. For this purpose, we designed a website at which users can enter a URL to differentiate between fake and legitimate URLs. The proposed model is composed of a 1D convolutional neural network (1D CNN). When tested on large-scale datasets from Phish Tank, UNB, and Alexa, the proposed model achieved good results and was tested on 200k phishing URLs and 200k legitimate URLs. The proposed model was compared with state-of-the-art architectures, and in the comparison, it achieved the highest accuracy of 99.7%

References:

1. B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, "Fighting against phishing attacks: state of the art and future challenges," Neural Computing and Applications, vol. 28, no. 12, pp. 3629–3654, 2017.
2. N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41, no. 13, pp. 5948–5959, 2014.
3. Apwg trends report," https://docs.apwg.org/reports/apwg trends report q1 2020.pdf.
4. L. Li, E. Berki, M. Helenius, and S. Ovaska, "Towards a contingency approach with whitelist-and blacklist-based anti-phishing applications: what do usability tests indicate?" Behaviour & Information Technology, vol. 33, no. 11, pp. 1136–1147, 2014.

5. B. Wardman, T. Stallings, G. Warner, and A. Skjellum, “High-performance content-based phishing attack detection,” in 2011 eCrime Researchers Summit. IEEE, 2011, pp. 1–9.
6. K. L. Chiew, E. H. Chang, W. K. Tiong et al., “Utilisation of website logo for phishing detection,” Computers & Security, vol. 54, pp. 16–26, 2015.
7. A. Alswailem, B. Alabdullah, N. Alrumayh, and A. Alsedrani, “Detecting phishing websites using machine learning,” in 2nd International Conference on Computer Applications & Information Security (ICCAIS). Riyadh, Saudi Arabia: IEEE, 2019, pp. 1–6.
8. M. Aydin and N. Baykal, “Feature extraction and classification phishing websites based on url,” in IEEE Conference on Communications and Network Security (CNS). Florence, Italy: IEEE, 2015, pp. 769–770.