Sarita Sahu, et. al. International Journal of Engineering Research and Applications
www.ijera.com
ISSN: 2248-9622, Vol. 14, Issue 12, December 2024, pp 80-84

RESEARCH ARTICLE OPEN ACCESS

Hybrid Machine Learning Technique for Prediction of Phishing Websites

Sarita Sahu[1],Onkar Nath Thakur[2], Rakesh Kumar Tiwari[3], Dr. Vikas Gupta[4]
M.Tech Scholar[1], Assistant Professor[2&3], Professor[4]
Department of Computer Science &Engineering[1,2&3],
Department of Electronics&Communication Engineering[4]
Technocrats Institute of Technology&Science, Bhopal, India

Date of Submission: 12-12-2024 Date of acceptance: 25-12-2024

Abstract- Phishing attacks represent a significant threat to online security, often resulting in substantial financial losses and data breaches. Despite the evolution of various security measures, the dynamic and sophisticated nature of phishing websites continues to outpace traditional detection methods. This paper presents a novel hybrid machine learning technique for the prediction of phishing websites, combining the strengths of multiple algorithms to enhance detection accuracy and robustness. The proposed approach integrates feature selection, ensemble learning, and deep learning models to create a comprehensive predictive framework. Through extensive experimentation on diverse datasets, the hybrid model demonstrates superior performance in identifying phishing websites compared to standalone machine learning models.

Index Terms- AI, Machine Learning, Hybrid, Phishing Websites.

I. INTRODUCTION

Phishing is a pervasive and evolving cyber threat that targets individuals and organizations by attempting to deceive them into disclosing sensitive information such as usernames, passwords, and credit card details [1]. These attacks are typically executed through deceptive emails and websites that mimic legitimate entities. The consequences of successful phishing attacks can be severe, ranging from financial loss to identity theft, and can undermine trust in online services. Given the increasing sophistication of phishing techniques, there is a pressing need for advanced detection mechanisms that can effectively identify and mitigate these threats [2].

Traditional methods for detecting phishing websites, such as blacklisting and heuristic-based approaches, have proven to be insufficient due to their limited adaptability and high false-positive rates [3]. Blacklisting, for example, cannot keep pace with the rapid creation of new phishing sites, while heuristic methods often fail to capture the nuances of modern phishing tactics [4]. Consequently, there is a growing interest in leveraging machine learning (ML) techniques to develop more robust and adaptive phishing detection systems.

Machine learning offers several advantages in phishing detection, including the ability to learn from large datasets, identify complex patterns, and make predictions based on new, unseen data [5]. However, single ML algorithms often fall short in achieving high accuracy and low false-positive rates due to the inherent complexity and diversity of phishing attacks. To address these challenges, hybrid machine learning techniques have emerged as a promising solution. These techniques combine multiple algorithms to capitalize on their individual strengths and mitigate their weaknesses, resulting in more accurate and reliable detection systems[6].

In this paper, we propose a hybrid machine learning technique for the prediction of phishing websites, integrating feature selection, ensemble learning, and deep learning models. Feature selection is employed to identify the most relevant attributes that contribute to distinguishing phishing sites from legitimate ones [7]. Ensemble learning, which combines the predictions of multiple base learners, is utilized to improve the overall predictive performance and robustness of the model. Additionally, deep learning models are incorporated to capture intricate patterns and relationships within the data that traditional algorithms might overlook [8].

The proposed hybrid approach is evaluated using extensive experiments on diverse phishing datasets. We compare the performance of the hybrid model against several standalone machine learning algorithms, including decision trees, support vector machines, and neural networks [9]. The results

demonstrate that the hybrid model outperforms these individual models in terms of accuracy, precision, recall, and F1 score. Furthermore, the hybrid technique significantly reduces the incidence of false positives, enhancing the reliability of phishing detection [10].

Research contributes to the field of cybersecurity by presenting an innovative hybrid machine learning approach for phishing website prediction. By leveraging the complementary strengths of various algorithms, the proposed model offers a robust and effective solution to combat the ever-evolving threat of phishing attacks, thereby enhancing the security and trustworthiness of online interactions.

The four sections of this study are as follows. The first part of this paper gives a general introduction to phishing detection. The suggested approach is presented in Section II, the simulation and results are presented in Section III, and a summary and conclusion are presented in Section IV.

II. PROPOSED METHODOLOGY

The following flowchart explains the suggested methodology:-

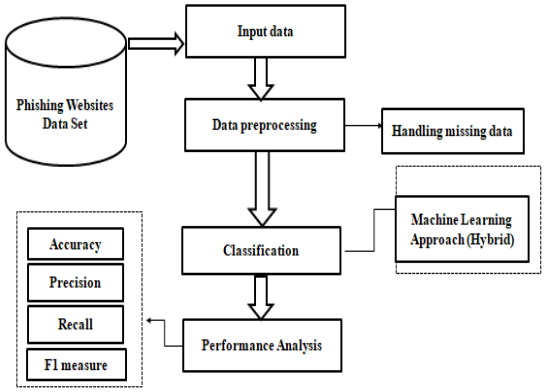

Figure 1: Flow Chart

Steps-

- Firstly, In the first place, complete the dataset [13] derived from the phishing website with data from a publicly accessible, huge dataset source.
- The data has been preprocessed, and the missing dataset is being sent over right now. Get rid of the blank spot by replacing it with a 1 or 0.
- Next, use a classification approach that takes use of the advantages of both traditional machine learning and the more recent hybrid approach.
- Precision, recall, F-measure, accuracy, and error rate are some of the performance metrics you should now examine and compute.

The suggested research technique relies on the following sub modules:

- Choosing and importing data
- Processing Information Ahead of Time
- Dataset Segmentation: Separating Test and Production Data
- Extraction of Features
- Classification
- Prediction
- Making an Impact

Data Selection and Loading

- The data selections are the process of selecting the dataset and load this dataset into the python environment.

Data Pre-processing

- Data selection is the process of choosing a dataset and importing it into a Python environment.
- The First Steps in Processing Data
- Unwanted information is filtered out of a dataset at the pre-processing stage.

Splitting Dataset into Train and Test Data

- Data splitting is the process of dividing a dataset into two halves, often for use in a cross-validator.
- The data is split in half, with one half used to create a prediction model and the other half for testing how well that model performed.

Feature Extraction

Data independence may be standardised via the use of feature extraction. It's often done in the pre-processing phase of data analysis and is also known as normalisation.

Classification- When it came to making the distinctions, a mixture of random forest and a gradient boosting classifier was used.

Random Forest-The first step of Random Forest is to generate an N-by-1 random forest of decision trees, and the second is to use those trees to produce predictions.

Detailed below are stages and a graphic explaining how the procedure works:

- Pick K training data points at random.
- Second, construct the decision trees connected to the points in question (Subsets).
- Third, decide on N as the total number of decision trees you'll be constructing.
- To recap, do steps 1 and 2 again in Step 4.

Find the forecasts of each decision tree for the new data points, and place them in the most popularly voted category.

Gradient Boosting- Gradient As one of the most popular optimization techniques, Gradient Boosting-Gradient Descent is widely used to train machine learning models by reducing the variance between observed and desired outcomes. The training of Neural Networks also makes use of gradient descent. Assumptions are made using starting parameters, and the cost function is computed by iteratively adjusting the parameters in the hope of lowering the cost function using gradient descent methods applied to previously collected data.

The definition of this term is the size of the steps required to get to the bottom. This is a very little number that is monitored and adjusted according on how the cost function is behaving. There is a trade-off between a faster learning rate and the potential for overshooting the minimum, with the former resulting in greater steps. However, a poor learning rate reveals the short step sizes, which trades overall efficiency for the sake of greater accuracy.

Prediction

- It's a method of spotting malicious apps for Android devices in a database.
- The overall performance of the research's prediction findings has been improved thanks to this study, and as a consequence, the data from the dataset has been accurately forecasted.

Result Generation

As a whole, categorization and prediction will be used to construct the final output. Different metrics, such as accuracy, error rate, etc., are used to assess how well the given method performs.

III. SIMULATION AND RESULTS

The Python Spyder IDE 3.7 is used for the simulation.

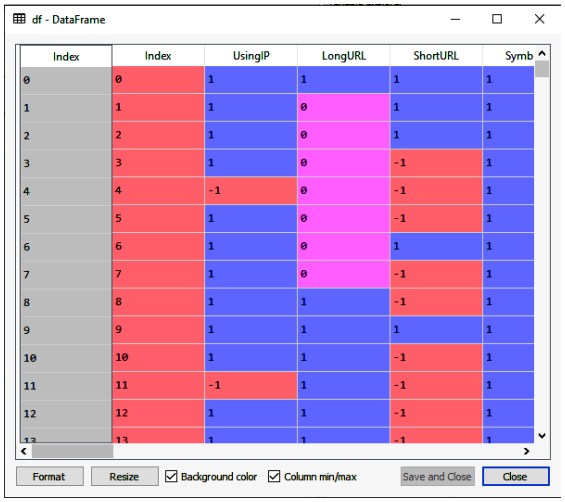

Figure 3: Dataset

It is shown in Figure 3 how the dataset looks in a python context. There is a wide range of row and column counts in the dataset. In each table, we identify the characteristics by name.

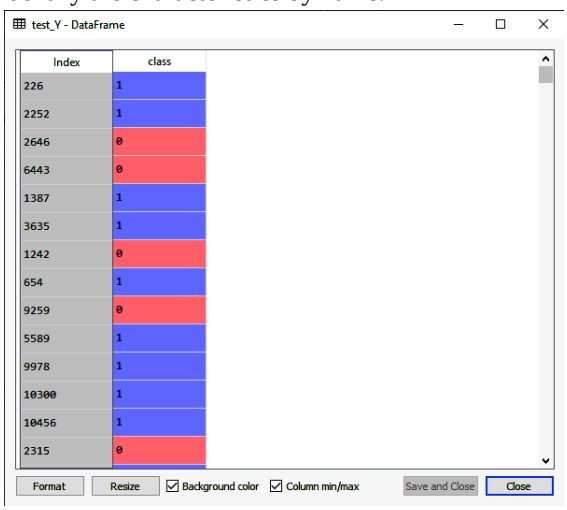

Figure 4: Y test

The y-test for the provided data is shown in Figure 4. Twenty percent to thirty percent of the original data set is used as the train data.

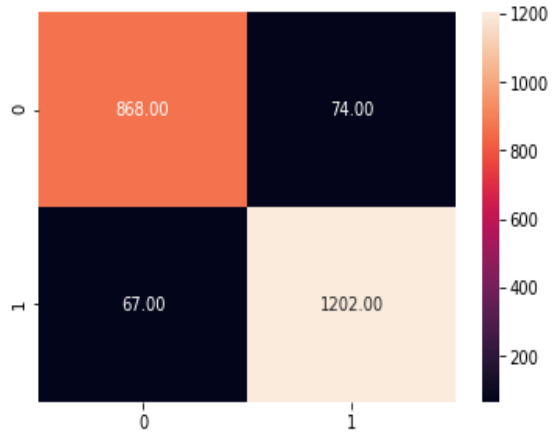

Figure 5: Confusion matrix heat map

The Hybrid classification technique's confusion matrix is shown as a heat map in Figure 5. To measure the efficacy of a classifier, statisticians utilise this N by N matrix.

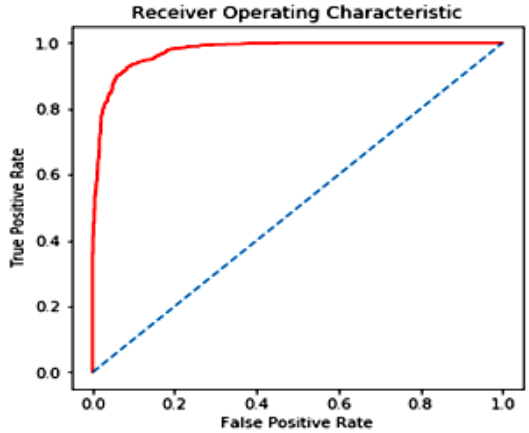

Figure 6: ROC

The ROC curve is shown in Fig. 6. Comparison of sensitivity (or TPR) and specificity may be seen in the ROC curve (1 – FPR). Better performance is represented by classifiers whose output curves are shifted to the upper left.

Table 1: Simulation Result

Sr. No.	Parameters	Value (%)
1	Accuracy	98.54%
2	Precision	98.60%
3	Recall	97.90%
4	F_Measure	98.3%
5	Classification error	1.46%

Table 2: Result Comparison

Sr. No.	Parameters	Previous Work	Proposed Work
1	Method	boosting based multi layer	Hybrid
2	Accuracy	96.79%	98.54%
3	Precision	96.84%	98.60%
4	Recall	96.70%	97.90%
5	F_Measure	96.77%	98.3%
6	Classification error	3.21%	1.46%

IV. CONCLUSION

The proposed hybrid machine learning technique for the prediction of phishing websites demonstrates a substantial improvement over traditional methods and previous approaches. The proposed hybrid machine learning technique significantly outperforms the previous boosting-based multi-layer approach across all key performance metrics. The hybrid model achieves an accuracy of 98.54%, a precision of 98.60%, and a recall of 97.90%, demonstrating its superior ability to correctly identify phishing websites. Additionally, the F-measure, which considers both precision and recall, is improved to 98.30%. The reduction in classification error to 1.46% further underscores the robustness and reliability of the proposed hybrid approach, making it a highly effective solution for phishing website detection.

REFERENCES

[1]. L. R. Kalabarige, R. S. Rao, A. R. Pais and L. A. Gabralla, "A Boosting-Based Hybrid Feature Selection and Multi-Layer Stacked Ensemble Learning Model to Detect Phishing Websites," in *IEEE Access*, vol. 11, pp. 71180-71193, 2023, doi: 10.1109/ACCESS.2023.3293649.

[2]. Y. Sun, G. Liu, X. Han, W. Zuo and W. Liu, "FusionNet: An Effective Network Phishing Website Detection Framework Based on Multi-Modal Fusion," *2023 IEEE International Conference on High Performance Computing & Communications, Data Science & Systems, Smart City & Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*,

Melbourne, Australia, 2023, pp. 474-481, doi: 10.1109/HPCC-DSS.

[3]. S. Mittal, R. Agarwal, M. L. Saini and A. Kumar, "A Logistic Regression Approach for Detecting Phishing Websites," *2023 International Conference on Advances in Computation, Communication and Information Technology (ICAICCIT)*, Faridabad, India, 2023, pp. 76-81, doi: 10.1109/ICAICCIT60255.2023.10466221.

[4]. J. M. Lindamulage, M. L, Y. S.P.J, P. I.S.S. and J. Krishara, "Vision GNN Based Phishing Website Detection," *2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Chennai, India, 2023, pp. 1-7, doi: 10.1109/ICSES60034.2023.10465358.

[5]. R. Sultana, M. A. Rahman and M. Ibrahim Khan, "Hybrid Model Based Phishing Websites Detection Using Deep Learning Technique," *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ICCIT60459.2023.10441639.

[6]. M. A. Snober, A. Droos and Q. A. Al-Haija, "Prevention of phishing website attacks in online banking systems using visual cryptography," *6th Smart Cities Symposium (SCS 2022)*, Hybrid Conference, Bahrain, 2022, pp. 168-173, doi: 10.1049/icp.2023.0391.

[7]. P. Jaswal, S. Sharma, N. Bindra and C. R. Krishna, "Detection and Prevention of Phishing Attacks on Banking Website," *2022 International Conference on Futuristic Technologies (INCOFT)*, Belgaum, India, 2022, pp. 1-8, doi: 10.1109/INCOFT55651.2022.10094345.

[8]. D. Ito, Y. Takata and M. Kamizono, "Money Talks: Detection of Disposable Phishing Websites by Analyzing Its Building Costs," *2022 IEEE 4th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, Atlanta, GA, USA, 2022, pp. 97-106, doi: 10.1109/TPS-ISA56441.2022.00022.

[9]. M. M. Uddin, K. Arfatul Islam, M. Mamun, V. K. Tiwari and J. Park, "A Comparative Analysis of Machine Learning-Based Website Phishing Detection Using URL Information," *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, Chengdu, China, 2022, pp. 220-224, doi: 10.1109/PRAI55851.2022.9904055.

[10]. L. Shalini, S. S. Manvi, N. C. Gowda and K. N. Manasa, "Detection of Phishing Emails using Machine Learning and Deep Learning," *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, Coimbatore, India, 2022, pp. 1237-1243, doi: 10.1109/ICCES54183.2022.9835846.

[11]. https://www.kaggle.com/datasets/isatish/phishing-dataset-uci-ml-csv?select=uci-ml-phishing-dataset.csv.