Phishing Website Detection Using Machine Learning

[1] Ishan Agarwal, [2] Gautam Bhatyani, [3] Harshit Sethi, [4] Abhinav Gaba

[1,2,3,4] Computer Science Department [1] Vani Rastogi

Abstract : Phishing involves a deceptive online tactic where an individual pretends to be trustworthy, aiming to deceive others into revealing sensitive information. An exemplary instance is trolling, which has posed an ongoing difficulty. Thankfully, recent progress in identifying phishing attempts, particularly through the application of machine learning, has led to a decrease in such occurrences. This research delves into constructing and contrasting four models to gauge machine learning's effectiveness in spotting phishing websites. Moreover, we evaluate the top-performing model against established methods documented in published studies. These models make use of (SVMs), (DTs) & (RF) .Our results indicate that the Random Forest model stands out as the most accurate among the four methods, surpassing them in precision, accuracy, and overall performance.

Keywords: phishing detection; phishing domains; random forest machine learning; support vector machine; decision tree;

1. Introduction :

Phishing represents an unlawful online practice aiming to deceive users into sharing personal information like usernames, passwords, financial data, addresses, and social connections. This data is maliciously exploited, often for identity theft. Orchestrated by hackers, phishing attacks involve impersonating trusted entities using sophisticated social engineering techniques. A variant of this is the phishing domain, where these domains illicitly gather personal information by coercing or redirecting users to deceptive websites mimicking legitimate ones. Both approaches rely on users divulging personal data, resulting in security breaches when users input their details into these deceptive websites, allowing attackers to exploit it for identity theft.

Cybersecurity endeavors to strengthen online defenses against threats like phishing domains by enhancing digital services offered by financial and governmental organizations. However, as reliance on online services like shopping, banking, and bill payments grows, the risks posed by phishing attacks to global financial systems become more pronounced, necessitating robust online protection. The core objective of cybersecurity is safeguarding Internet-connected resources from cyber-attacks.

The increasing frequency and complexity of cyber threats complicate the task of identifying, evaluating, and managing these significant risk events. In 2016, over 51,000 distinct phishing websites were detected, costing global enterprises $9 billion. In 2016, the documented count of phishing attacks exceeded 1 million, marking a 65% rise from the preceding year and eroding confidence in online platforms. Phishing, among various web frauds, prominently utilizes imitation web pages and fraudulent URLs distributed through spam chats or social media to deceive users into revealing sensitive information.

Various strategies can combat phishing. Artificial Intelligence (AI) plays a substantial role in cybersecurity, aiding in detecting spam, phishing, and sophisticated attacks like spear phishing by leveraging historical datasets for analysis. This research assesses Machine Learning (ML) classification models for the identification of phishing domains, with the goal of improving detection accuracy by identifying the most efficient model among the four, predicting whether a webpage is legitimate or poses a potential phishing threat. Evaluating a phishing domain is complex, necessitating a comprehensive approach that quantitatively and qualitatively analyzes its causes and features to guide the model, particularly concerning consumer trust.

The subsequent parts of this document follow a structured layout, with Section 2 delving into recent studies regarding phishing attacks. Section 3 delineates the data and ML-based solutions employed. Section 4 delves into the suggested model architecture. Section 5 thoroughly analyzes results and evaluates the proposed model. Ultimately, Section 6 wraps up the document and suggests possible avenues for future research.

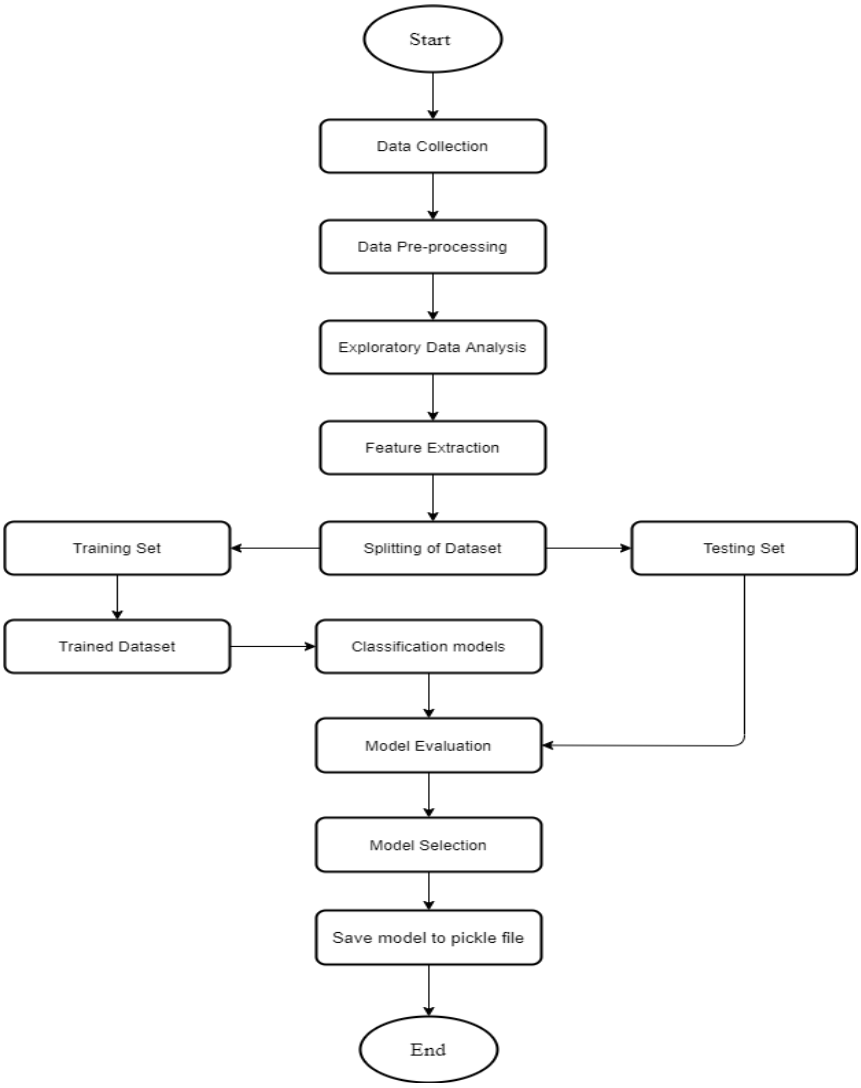


Fig 1.1 – Data Flow Diagram

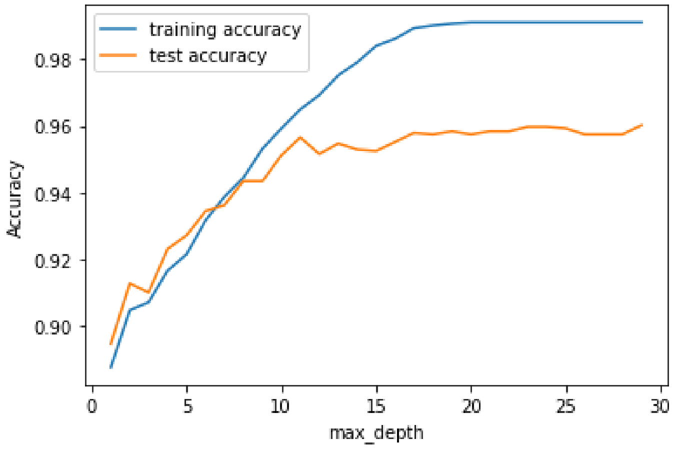
2. Background :

Commonly utilized classification methods in machine learning have demonstrated their capacity to adeptly identify and detect phishing domains. These approaches include:

2.1. Decision Tree

Decision trees improve decision-making by showcasing various options and their potential results in a structured tree layout. This method of modeling takes numerous factors into consideration, including potential outcomes, resource costs, and utility. In essence, a decision tree is an algorithm formed using conditional control statements, widely used for analyzing relationships within extensive datasets. Its primary purpose is to outline a process, enabling analysts to assign attributes to specific categories. The structure of a decision tree is developed through a training algorithm, allowing it to assess additional data samples with differing levels of accuracy, contingent on how well it captures the essence of the dataset. The rates of success are influenced by diverse factors, encompassing the dataset's size, intersections among variable observations, the choice of algorithm, and the integration of supplementary procedures to improve tree implementation.

Commencing from initial node, which signifies entire dataset, decision tree bifurcates into two consistent groups referred to as child nodes, ultimately reaching leaf nodes that depict the final output.Once the leaf node is reached, no additional divisions take place. The process of division entails breaking down the root node into sub-nodes according to particular conditions. The emergence of branches or sub-trees results from the division of a larger tree into smaller segments. In contrast, pruning involves eliminating unnecessary branches from the tree.



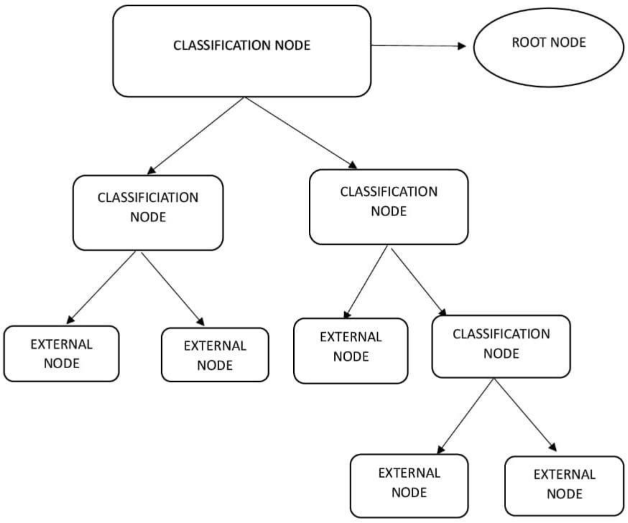
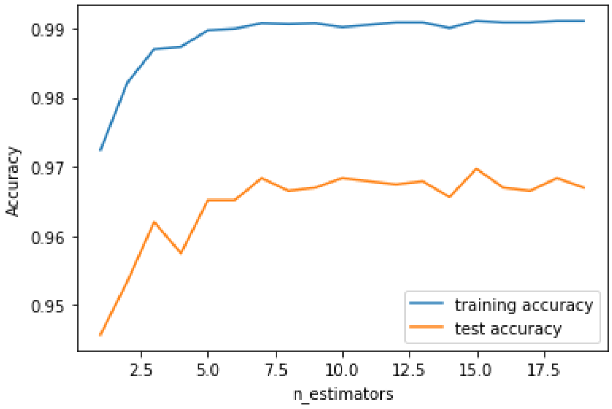


Fig 2.1 Decision Tree

2.2 Random Forest

The random forest is an amalgamation of guided decision tree algorithm employed in the field of machine learning for purposes such as classification and regression modeling. Its capability to rapidly and precisely categorize data has garnered significant attention in academic circles. This methodology combines outcomes and predictions from multiple decision trees to determine the most suitable output, which could either be the most prevalent class or the average prediction. To commence the procedure, the dataset is split into a training set and a test set, with multiple samples chosen in a random fashion. Each sample is then further divided into two branches using the most effective division method. This process is iterated to gather votes on each prediction, ultimately selecting the most commonly predicted outcome. Essential parameters of the random forest can be adjusted to enhance predictive ability or expedite model performance. While increasing the number of trees might boost prediction accuracy and stability, it can also elongate processing time. Enhancing algorithm performance involves using the highest acceptable number of features and the smallest permissible number of leaves.

The model can be employed at a test dataset to generate predictions, which are subsequently compared with the expected results. Each tree yields a distinct output when presented with a random sample vector. The random forest is employed to generalize errors, with its accuracy improving as the forest size expands. The accuracy, evaluated by randomly selecting features for error rate assessment, heavily relies on the interconnections among the trees. The properties of the random forest, encompassing error rates and node connections, ascertain the significance of a variable, offering insights into its importance.



2.3. Support Vector Machine

Support Vector Machines (SVM), based on principles from statistical learning theory, plays a dual role in both pattern recognition and regression. While this theory identifies fundamental components for mastering simple algorithms, real-world applications often demand more sophisticated tools like neural networks. However, the complexity inherent in these advanced tools poses challenges for theoretical analysis compared to practical scenarios. SVM adeptly reconciles this contrast by creating models that exhibit complexity akin to neural networks yet retain simplicity for mathematical examination. This simplicity stems from SVM's representation of a linear algorithm within a high-dimensional space.

Within the SVM framework, a hyperplane or decision boundary is delineated between two specified classes, each labelled at least once. Data points and support vectors are depicted by hyperplanes, and their distances are utilized to classify each class individually. Existing studies indicate that the hyperplane with the maximum margin between the two classes exhibits the most effective generalization performance. Support vectors, a subset of training instances, contain essential information for solving a classification problem, ensuring consistent results even if all other vectors are removed.

2.4. XGBOOST

SVM, based on statistical learning theory, fulfills a dual role in handling pattern recognition and regression. Although this theory identifies foundational aspects for mastering fundamental algorithms, practical scenarios often require more complex tools such as neural networks. However, the complexity inherent in these advanced tools presents challenges for theoretical analysis compared to real-world applications. SVM effectively addresses this by developing models that mimic the intricacy of neural networks while retaining simplicity for mathematical examination. This simplicity is

derived from SVM's portrayal of a linear algorithm within a high-dimensional space.

Within the SVM framework, a hyperplane or decision boundary delineates two specified classes, each labeled at least once. Data points and support vectors are seen as hyperplanes, and their distances are utilized for self-classification of each class. Prior studies suggest The hyperplane demonstrating the most substantial separation between the two classes achieves optimal generalization performance.

Determining the optimal hyperplane entails solving a convex optimization problem, which includes the minimization of a quadratic function under linear inequality constraints. Support vectors, which constitute a subset of training instances, are part of this process, contain all essential information for resolving a classification problem, ensuring consistent outcomes even if all other vectors are eliminated.

3. Methodology:

Four models were created employing ANN, SVM, DTs, and RF algorithms on the UCI dataset for the detection of phishing attacks. MinMax normalization served as a preprocessing technique to enhance the accuracy of these models. These models proficiently detected various forms of attacks within the UCI dataset. For additional details, comprehensive discussions regarding the dataset and implemented algorithms can be found in sections 4.1 and 4.2.

The UCI repository provides accessible existing datasets for developing algorithms aimed at spotting phishing sites. Certain studies focus on classifying websites to generate separate lists of legitimate and phishing sites for subsequent analysis. This research utilizes the freely accessible phishing dataset obtained from UCI's machine learning repository, which has been curated by [17].Tailored for building machine learning algorithms to identify phishing websites, this dataset includes a comprehensive set of attributes across four distinct categories [18]. The selected attributes were deliberately chosen and extracted from aspects such as Address Bar, HTML/JavaScript, Unusual Behavior, and Domain. The study employed a dataset specific to domains, featuring 31 attributes, each with binary or ternary values. This dataset includes a total of 11,055 entries, and each entry incorporates 31 features like URL length, email submission, using a shortening service, unusual URL structures, presence of an "@" symbol, redirection, and other factors.

To improve accuracy, MinMax standardization was employed in each model in this study. Normalization plays a vital role in enhancing the precision of ML models and is essential for the effective operation of certain models. MinMax standardization compresses data to a range of [0, 1], enhancing the quality of input data for model training. Equations (1) and (2) demonstrate the MinMax standardization process, where

$x\_\_std = (x - x.min) / (x.max - x.min)$

$x\_scale = x\_\_std \times (xmax - xmin)$

To enhance efficiency and address intricacies, the algorithm employed a data normalization technique outlined in Table 2. The crucial elements are extracted from the original dataset by evaluating the predicted outcome based on 30 features. The UCI dataset was then partitioned into training sets (80/20) and testing sets (50/50) using c5-fold cross-validation, a method acknowledged for optimal performance in recent studies. The utilization of diverse learning models in machine learning guides the prediction model, preventing bias toward any single model, which is considered a beneficial practice. Summarizing the models and computing their highest accuracies, if a domain is identified as phishing by most models, The accuracy of the model's predictions reflects the probability that the domain is a phishing attempt.

Table 1: Illustrates the outcomes of the performance both prior to and following the application of the normalization method.

Classifiers	Before Normalization	After Normalization
SVM	Accuracy:93.89 Precision: 94.49 Recall: 95.87 F1-measure: 96.01	Accuracy: 95.08 Precision: 94.1 Recall: 95.98 F1-measure: 96.21
XGBOOST	Accuracy: 96.74 Precision: 97.96 Recall: 96.87 F1-measure: 96.80	Accuracy: 97.68 Precision: 97.76 Recall: 97.69 F1-measure: 98.21
RF	Accuracy:95.69 Precision: 95.89 Recall: 98.01 F1-measure: 96.59	Accuracy: 96.98 Precision: 96.26 Recall: 98.36 F1-measure: 98.25
DT	Accuracy: 95.36 Precision: 96.11 Recall: 96.12 F1-measure: 94.82	Accuracy: 97.02 Precision: 96.63 Recall: 97.32 F1-measure: 95.2

4. Results and Findings

This research employed an experimental methodology to recognise the ML models that is most efficient for detecting phishing domains. Four machine learning methods—SVM, XGBOOST, RF, and DT—were employed in this investigation. The UCI dataset, comprising 11,055 data instances, formed the cornerstone of this research. The assessment of this dataset involved thirty features, with the 31st feature designated as the resultThe results of the simulation, as outlined in Table 3, encompassed metrics like the true positive rate, false positive rate, true negative rate, and false negative rate for assessment.To improve accuracy of the classification process, a 5x5x Cross-validation approach was implemented during the experimentation.

A technique employed to assess predictive performance models by gauging their predictive abilities on external data. Assessing the confusion matrix is vital for evaluating how effective the classification technique is. The findings in Table 4 show that XGBOOST had the highest detection accuracy, with DT at 95%, RF at 96%, and SVM following in accuracy.

Table 2. Results of the evaluation presented in percentage terms(%).

Classifiers Accuracy	Precision	Recall	F1- Measure
SVM 0.964	0.965	0.98	0.968
XGBOOST 0.968	0.991	0.993	0.972
RF 0.967	0.99	0.992	0.967
DT	0.96	0.993	0.991 0.964

Additional studies exploring phishing attacks and diverse machine learning approaches are discussed in the research conducted by Ubing et al. [19]. They created three ensemble learning methods—bagging, boosting, and stacking—achieving a 95.4% success rate through the amalgamation of their classifiers. Lakshmi et al. [20] presented an imaginative method for detecting phishing websites, scrutinizing HTML pages in related websites' source code for links, resulting in a 96% success rate. In another study [20], researchers proposed three meta-learner models using ForestPA. Their experimental data demonstrated accuracy ranging from 95% to 97%, with the lowest accuracy recorded at 97.4%, except for Alsariera et al. [21], who achieved the same 97.4% accuracy rate. However, it's worth noting that this model demands more time for training and implementation compared to RF or DT classifiers. Cross-validation serves as an evaluation.

5. Conclusion

This study explored the feasibility and productivity of implementing machine learning for the recognision of phishing.Four ML models, namely XGBoost, SVM, DTs, and RF were created. Research identified the most robust model among these and conducted a comparison with established solutions found in the existing literature. The results showed that the XGBoost (XGB) model exhibited outstanding performance, outstripping other methods documented in prior studies. Future investigations will focus on exploring additional machine learning algorithms for identifying phishing domains.

References

[1] Cabaj, K.; Domingos, D.; Kotulski, Z.; Respício, A. Cybersecurity Education: Evolution of the Discipline and Analysis of Master Programs. Comput. Secur. 2018, 75, 24–35. [CrossRef]

[2] Iwendi, C.; Jalil, Z.; Javed, A.R.; Reddy, G.T.; Kaluri, R.; Srivastava, G.; Jo, O. KeySplitWatermark: Zero Watermarking Algorithm for Software Protection Against Cyber-Attacks. IEEE Access 2020, 8, 72650–72660. [CrossRef]

[3] Rehman Javed, A.; Jalil, Z.; Atif Moqurrab, S.; Abbas, S.; Liu, X. Ensemble Adaboost Classifier for Accurate and Fast Detection of Botnet Attacks in Connected Vehicles. Trans. Emerg. Telecommun. Technol. 2020, 33, e4088. [CrossRef]

[4] Conklin, W.A.; Cline, R.E.; Roosa, T. Re-Engineering Cybersecurity Education in the US: An Analysis of the Critical Factors. In Proceedings of the 2014 47th Hawaii International Conference on System Sciences, IEEE, Waikoloa, HI, USA, 6–9 January 2014; pp. 2006–2014.

[5] Javed, A.R.; Usman, M.; Rehman, S.U.; Khan, M.U.; Haghighi, M.S. Anomaly Detection in Automated Vehicles Using Multistage Attention-Based Convolutional Neural Network. IEEE Trans. Intell. Transp. Syst. 2021, 22, 4291–4300. [CrossRef]

[6] Bleau, H.; Global Fraud and Cybercrime Forecast. Retrieved RSA 2017. Available online: https://www.rsa.com/en-us/resources/ 2017global-fraud (accessed on 19 November 2021).