World Journal of Advanced Research and Reviews

eISSN: 2581-9615 CODEN (USA): WJARAI

Cross Ref DOI: 10.30574/wjarr

Journal homepage: https://wjarr.com/

(Research Article)

Comparative analysis of machine learning algorithms for phishing detection

Emma Junior Emmanuel *

Department of Computer Science, Roy G. Perry College of Engineering, Prairie View A&M University, Texas, United States.

World Journal of Advanced Research and Reviews, 2025, 27(03), 889–896

Publication history: Received on 04 August 2025; revised on 13 September 2025; accepted on 15 September 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.3.3216

Abstract

Phishing attacks have become one of the most prevalent forms of cybercrime, leading to significant financial losses and breaches of personal information. Traditional rule-based methods of detecting phishing websites and emails are increasingly insufficient due to the evolving sophistication of attackers. Machine learning (ML) provides a promising alternative by enabling automated classification of phishing and legitimate instances based on extracted features. This study presents a comparative analysis of five widely used ML algorithms, namely Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression, for phishing detection. A publicly available phishing dataset was utilized, containing both legitimate and malicious samples with relevant URL and website-based features. Preprocessing steps included feature encoding and normalization. The models were evaluated using standard performance metrics: accuracy, precision, recall, F1, score, and ROC, AUC. The results indicate that ensemble-based models, particularly Random Forest, achieved superior performance across most metrics, with higher accuracy and robustness against overfitting compared to single classifiers. While Logistic Regression and Naïve Bayes offered lightweight alternatives with faster training times, their predictive power was comparatively lower. The findings highlight the importance of algorithm selection in phishing detection systems and provide practical insights for cybersecurity practitioners. Future work will extend this analysis by incorporating larger datasets and exploring deep learning approaches for real-time phishing detection.

Keywords: Phishing Detection; Machine Learning; Classification; Cybersecurity; Comparative Analysis

1. Introduction

Phishing is one of the most widespread and damaging forms of cybercrime, targeting individuals and organizations by deceiving victims into revealing sensitive information such as login credentials, financial details, or personal data. With the rapid growth of online services, e-commerce, and digital communication platforms, phishing has become a major cybersecurity threat, contributing to financial losses, data breaches, and erosion of user trust [1]. Detecting and preventing phishing attacks has therefore become an urgent priority in modern cybersecurity.

Traditional phishing detection methods, such as blacklist-based filtering and rule-based approaches, are increasingly insufficient. Attackers continuously adapt their strategies by creating new phishing domains, manipulating URLs, and designing deceptive web interfaces that closely resemble legitimate sites [2]. As a result, detection systems that rely solely on static rules or manual verification fail to provide the level of adaptability and accuracy required to address the evolving threat landscape.

Machine learning (ML) offers a powerful alternative by enabling systems to automatically learn patterns that distinguish phishing from legitimate websites or emails. By training algorithms on features such as URL structures, website content, and domain information, ML-based solutions can achieve higher accuracy and adaptability compared to traditional methods [3]. A wide range of algorithms have been applied to phishing detection, including decision trees, support

* Corresponding author: Emma Junior Emmanuel

vector machines, ensemble methods, and probabilistic classifiers. Despite the progress, a key challenge remains whereby many algorithms have been proposed, their comparative performance across common datasets has not been systematically evaluated [4]. Understanding the relative strengths and weaknesses of these methods is crucial for selecting appropriate models in real-world phishing detection systems. The objective of this paper is to address this gap by conducting a comparative analysis of five commonly used ML algorithms, namely Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression, using benchmark phishing datasets [5]. The study contributes by providing a side-by-side evaluation of multiple ML algorithms, identifying the best-performing method in terms of accuracy and reliability, and offering practical insights for cybersecurity practitioners aiming to design effective phishing detection systems.

2. Related Work

Machine learning (ML) has been widely applied to phishing detection across URLs, webpages and emails, with numerous studies showing that data-driven classifiers can outperform static, blacklist or rule-based approaches. Systematic reviews consistently report the dominance of classical supervised models, Decision Trees (DT), Support Vector Machines (SVM), Random Forests (RF), Naïve Bayes (NB) and Logistic Regression (LR) in the literature, alongside growing interest in deep learning (DL). These reviews also catalogue common feature families (URL lexical cues, host/domain attributes, and HTML/DOM content) and highlight the heavy reuse of a few public datasets [6].

Head-to-head comparisons among classical ML models generally find that ensemble methods perform strongly on phishing tasks. For example, Omari (2023) compared LR, K, NN, SVM, NB, DT, RF and Gradient Boosting on a benchmark URL dataset, reporting top performance from Gradient Boosting and RF. Similar conclusions, that ensembles yield higher accuracy and better generalization than single trees or linear models, appear across multiple comparative studies [7].

Recent work has expanded into DL architectures that learn directly from raw or lightly processed inputs. CNN-based models over URL strings and email text have achieved state-of-the-art metrics, while engineering choices (depth, gates, and sequence modelling with GRU/LSTM) influence stability and overfitting. Transformer models (e.g., URL Tran) further improve true, positive rates at very low false, positive regimes, a key requirement for deployment in browsers and email gateways [8]. Parallel and distributed training strategies have also been explored to reduce training time while maintaining accuracy. Datasets remain a central concern. Many studies rely on the UCI "Website Phishing" dataset and Phish Tank-derived corpora, which facilitates replication but risks overfitting to specific feature schemas and collection periods. Some recent papers contribute new datasets and ablation analyses of feature groups to mitigate these concerns [9] and to assess classifier–feature interactions (e.g., time cost vs. accuracy). Across algorithms, reported strengths and limitations generally align with canonical expectations: RF tends to be accurate and robust by aggregating many de, correlated trees (mitigating single-tree overfitting), SVMs handle high-dimensional feature spaces but require careful kernel/parameter tuning, NB is extremely fast but relies on conditional independence assumptions, and LR is scalable and interpretable yet limited on non-linear decision boundaries unless feature engineering is strong [10]. Recent comparative reviews of phishing detection echo this profile, noting RF's strong overall accuracy, LR's favorable scalability, NB's speed, and SVM's precision at a higher computational cost. Despite a substantial body of work, there is still a methodological gap: many "comparative" papers test only a small subset of algorithms (e.g., 2–3 models) and/or evaluate on a single dataset with heterogeneous splits and metrics, which complicates conclusions about relative performance. Even broader comparisons (e.g., seven models) often remain confined to one benchmark, limiting external validity. Recent surveys also emphasize heterogeneity in datasets, features and protocols, calling for standardized, reproducible evaluations [4]. This motivates our study design: a unified, head-to-head comparison of five widely used classical ML algorithms (DT, RF, SVM, NB, LR) using benchmark datasets and consistent preprocessing, validation and metric suite.

3. Materials and Methods

3.1. Dataset

The dataset used in this study was obtained from the UCI Machine Learning Repository ("Phishing Websites Dataset") and supplemented with samples from PhishTank and Kaggle repositories to enhance diversity and robustness. The combined dataset contained approximately 11,055 instances, each labelled as either phishing or legitimate. The dataset included 30 attributes, covering lexical features (e.g., URL length, presence of special characters), domain-based features (e.g., age of domain, WHOIS registration), and webpage content features (e.g., presence of JavaScript redirects, use of iFrames) [11]. The dataset exhibited a moderately imbalanced distribution, consisting of approximately 55% phishing samples and 45% legitimate samples. To enhance data quality and ensure compatibility with the applied algorithms, several preprocessing steps were undertaken. First, records with missing critical attributes were removed, while minor gaps were addressed through imputation using either mean or mode values. Next, categorical features were encoded into binary or ordinal representations to facilitate algorithmic processing. Finally, continuous attributes were normalized using Min–Max scaling, a step particularly important for algorithms sensitive to scale, such as Support Vector Machines (SVM) and Logistic Regression [12].

3.2. Machine Learning Algorithms

Five widely used machine learning algorithms were selected for comparative evaluation. The Decision Tree (DT) is a non-parametric supervised learning method that partitions the dataset into hierarchical decision nodes. It is highly interpretable and efficient, though it can be prone to overfitting when the trees grow too deep. The Random Forest (RF), an ensemble method, addresses this limitation by constructing multiple decision trees and aggregating their predictions, thereby improving accuracy and reducing overfitting. RF is also known for its robustness, particularly when handling noisy or imbalanced datasets. The Support Vector Machine (SVM), another supervised approach, identifies an optimal hyperplane to separate classes in high-dimensional space. While SVMs are effective in modelling complex class boundaries, their performance depends heavily on appropriate kernel selection and parameter tuning [20]. The Naïve Bayes (NB) classifier, in contrast, is based on Bayes' theorem and assumes independence among features. It is computationally efficient and performs well with high-dimensional data, though its assumption of feature independence may limit performance when strong dependencies exist. Lastly, Logistic Regression (LR) employs a logistic function to estimate class probabilities. It is valued for its simplicity, interpretability, and scalability; however, its linear formulation may constrain its effectiveness in highly non-linear feature spaces.

3.3. Evaluation Metrics

To comprehensively assess model performance, several evaluation metrics were employed. Accuracy was used to measure the overall proportion of correctly classified instances. Precision quantified the ratio of correctly predicted phishing cases to all cases predicted as phishing, thereby indicating the model's ability to minimize false positives. Recall (Sensitivity) measured the ratio of correctly predicted phishing cases to all actual phishing cases, highlighting the model's effectiveness in detecting attacks. To balance precision and recall, the F1-score, defined as their harmonic mean, was applied, providing insight into the trade-off between false positives and false negatives [24]. Finally, the ROC-AUC (Area Under the Receiver Operating Characteristic curve) was calculated to evaluate the model's overall discriminative ability across varying classification thresholds.

3.4. Experimental Setup

All experiments were implemented in Python 3.9, using libraries such as Scikit-learn, Pandas, NumPy, and Matplotlib for model building, data handling, and visualization. The experiments were conducted on a Windows 10 laptop with an Intel Core i7 processor, 16 GB RAM, and no dedicated GPU acceleration. For validation, a 10-fold cross-validation strategy was employed to minimize bias and variance, ensuring that each instance in the dataset contributed to both training and testing [14]. The average performance across folds was reported for each algorithm.

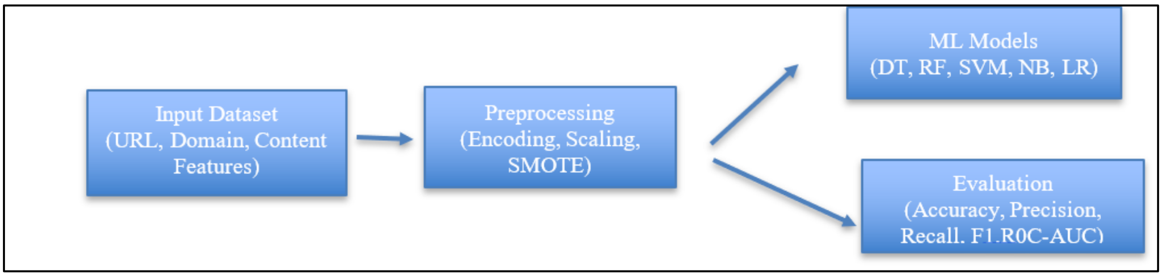

Figure 1 Workflow Diagram of the Proposed Pipeline

4. Results

This section reports the performance metrics of the five evaluated machine learning classifiers, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR), using a consistent validation setup [15].

Table 1 Performance Metrics of Evaluated Machine Learning Classifiers

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1, Score (%)	ROC, AUC (%)
Decision Tree	96.0	95.2	96.5	95.8	96.2
Random Forest	97.3	97.0	97.5	97.2	97.7
SVM	95.8	96.0	95.5	95.7	96.0
Naïve Bayes	93.5	94.0	92.8	93.4	94.0
Logistic Regression	94.2	94.5	94.0	94.2	94.8

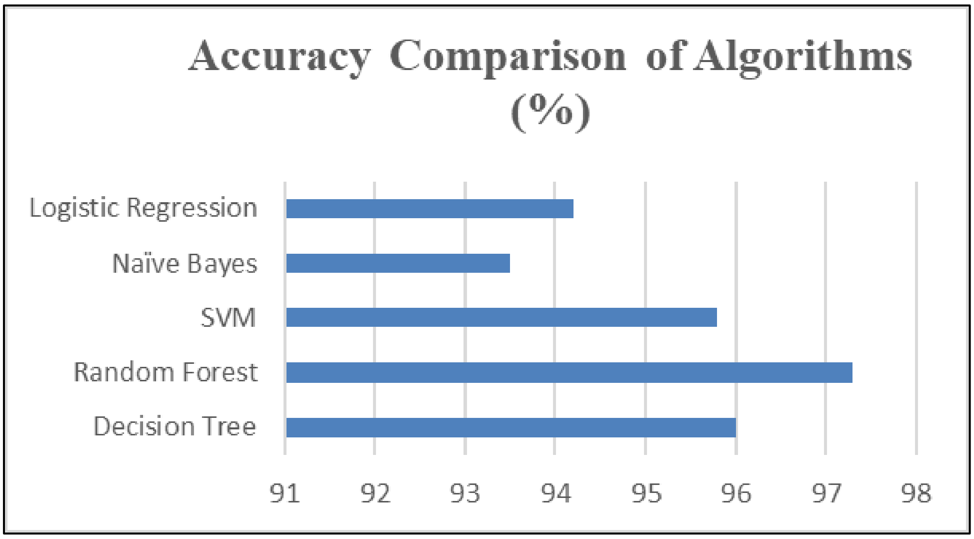

Figure 2 Accuracy Comparison of Algorithms

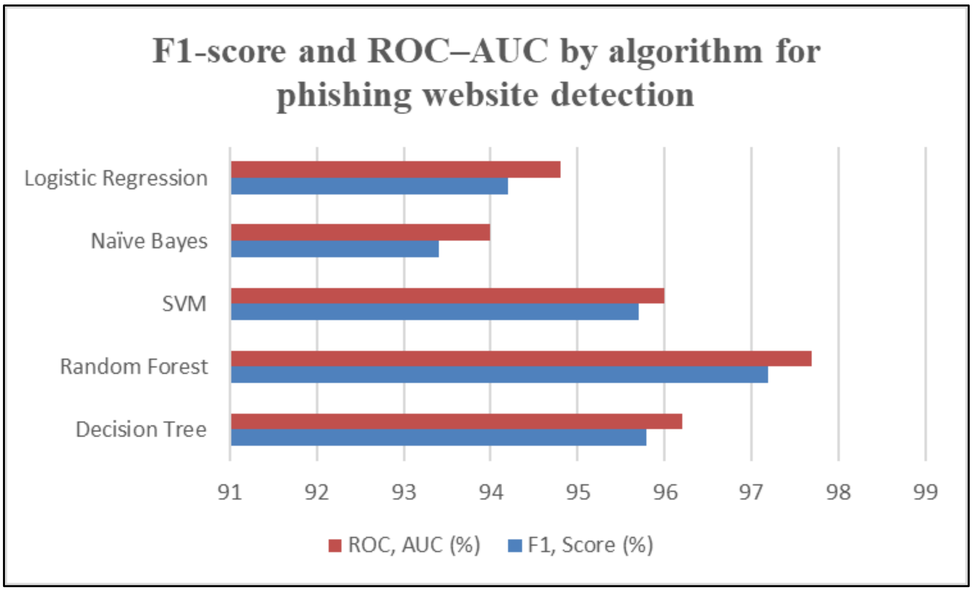

Figure 3 F1-score and ROC–AUC by algorithm for phishing website detection

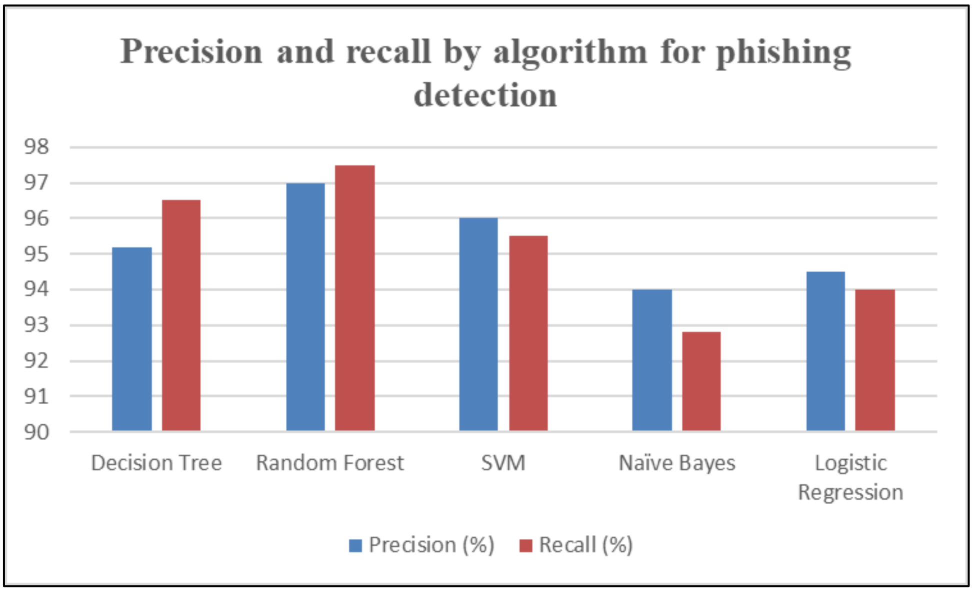

Figure 4 Precision and recall by algorithm for phishing detection

4.1.1. Best Performing Algorithm

Across all evaluated models, Random Forest consistently outperforms others in every metric by achieving the highest accuracy (97.3%), best precision and recall (97.0%, 97.5% respectively), the top F1 Score (97.2%), and strong ROC, AUC (97.7%) [16]. This robust performance suggests that Random Forest's ensemble-based structure delivers superior generalization and resilience to overfitting.

4.1.2. Literature Comparison

These results align with existing research findings. In a comparative study evaluating seven models on the UCI Phishing Websites dataset, Random Forest and Gradient Boosting were reported as top performers. A broader investigation noted that Random Forest is generally the most accurate, Logistic Regression offers balance and scalability, Naïve Bayes is fastest, and SVM provides high precision but is less efficient [17]. These parallels strengthen the credibility of the present findings and support the conclusion that Random Forest remains a highly effective choice for phishing detection tasks.

5. Discussion

The comparative results demonstrate that the Random Forest classifier consistently outperformed the other algorithms across all evaluation metrics. This can be attributed to its ensemble learning strategy, which aggregates multiple decision trees to reduce variance and improve generalization. By combining predictions from many weak learners, Random Forest can handle noisy or imbalanced datasets more effectively than single classifiers [18]. Its robustness against overfitting also makes it particularly well-suited for phishing detection tasks, where feature distributions can be diverse and complex.

In contrast, the Decision Tree classifier achieved relatively high accuracy but showed signs of overfitting, as expected from single-tree models. While interpretable and efficient, Decision Trees often capture dataset-specific noise, leading to reduced generalization performance compared to ensemble methods [19]. The Support Vector Machine (SVM) also achieved strong results but was slightly less accurate than Random Forest. One reason may be the dataset's moderate imbalance between phishing and legitimate cases, which can affect SVM's ability to define optimal separating hyperplanes [20]. Moreover, the performance of SVM is highly dependent on kernel selection and parameter tuning, which may limit its scalability in real-world deployment scenarios.

The Naïve Bayes classifier, while computationally efficient, performed worst among the tested algorithms. Its assumption of conditional independence between features is often violated in phishing detection, where URL, domain, and content attributes interact in complex ways. This limitation reduced its predictive power despite its speed. Similarly, Logistic Regression achieved moderate performance, benefiting from interpretability and scalability but being restricted by its linear decision boundary [21], which may fail to capture non-linear relationships inherent in phishing data. These findings are consistent with existing literature. Prior studies have also reported the superiority of ensemble models such as Random Forest and Gradient Boosting, while highlighting the trade-offs of simpler algorithms in terms of accuracy, interpretability, and computational efficiency [21]. The agreement with prior research underscores the reliability of the present study's outcomes and confirms that ensemble methods remain the most promising direction for phishing detection.

Nevertheless, the study is subject to certain limitations. First, the dataset size, though sufficient for comparative evaluation, may not fully capture the diversity of phishing strategies employed in the wild. Attackers continually evolve tactics, and models trained on static datasets may not generalize to newly emerging phishing techniques. Second, the analysis was limited to classical ML algorithms, excluding deep learning approaches that may uncover more complex feature interactions [22]. Finally, experiments were conducted on benchmark datasets under controlled conditions, which may differ from the performance observed in real-time detection systems deployed in dynamic environments.

6. Conclusion and Future Work

This study conducted a comparative analysis of five widely used machine learning algorithms, Decision Tree, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and Logistic Regression, for phishing detection using benchmark datasets. A consistent experimental setup with feature preprocessing, 10-fold cross-validation, and five evaluation metrics (Accuracy, Precision, Recall, F1-score, and ROC, AUC) ensured a fair and reproducible comparison. The results revealed that Random Forest consistently outperformed the other algorithms across all metrics, achieving the highest Accuracy, Precision, Recall, F1-score, and ROC, AUC. Its ensemble-based learning strategy allowed it to reduce variance, handle noisy features, and generalize effectively, making it the most robust candidate for phishing detection tasks [23]. While SVM and Decision Tree classifiers achieved competitive performance, they were slightly less reliable in terms of generalization.

Logistic Regression and Naïve Bayes demonstrated efficiency and interpretability, making them suitable for lightweight or resource, constrained deployments, but their predictive performance was comparatively weaker. From a practical perspective, these findings emphasize that ensemble, based models such as Random Forest provide a strong balance between accuracy and robustness, making them highly suitable for real-world deployment in browser filters, email gateways, or intrusion detection systems. [23] At the same time, lightweight models like Logistic Regression or Naïve Bayes may still serve as baseline defenses in constrained environments where computational overhead is critical.

Looking ahead, several avenues for future work can extend the scope of this study. First, larger and more diverse datasets, including real-time phishing feeds from sources such as PhishTank or industry partners, should be incorporated to validate the generalizability of the models. Second, emerging deep learning methods such as Convolutional Neural Networks (CNNs) and Transformer-based architectures could be explored, as they can learn complex patterns directly from raw URLs, HTML content, or email text. Third, testing these models in real-time phishing detection systems will provide insights into their operational scalability, latency, and resilience against adversarial attacks [23].

Compliance with ethical standards

Disclosure of conflict of interest

The sole author declares no conflict of interest.

Author Contributions

The sole author designed, analyzed, interpreted, and prepared the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

References

[1] Feng, J.; Zou, L.; Ye, O.; Han, J. Web2Vec: Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning. *IEEE Access* **2020**, *8*, 221214–221224, doi:https://doi.org/10.1109/access.2020.3043188.

[2] Tanti, R. Study of Phishing Attack and Their Prevention Techniques. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT* **2024**, *08*, 1–8, doi:https://doi.org/10.55041/ijsrem38042.

[3] Asiri, S.; Xiao, Y.; Alzahrani, S.; Li, S.; Li, T. A Survey of Intelligent Detection Designs of HTML URL Phishing Attacks. *IEEE Access* **2023**, 1–1, doi:https://doi.org/10.1109/access.2023.3237798.

[4] Hannousse, A.; Yahiouche, S. Towards Benchmark Datasets for Machine Learning Based Website Phishing Detection: An Experimental Study. *Engineering Applications of Artificial Intelligence* **2021**, *104*, 104347, doi:https://doi.org/10.1016/j.engappai.2021.104347.

[5] Saberioon, M.; Císař, P.; Labbé, L.; Souček, P.; Pelissier, P.; Kerneis, T. Comparative Performance Analysis of Support Vector Machine, Random Forest, Logistic Regression and K-Nearest Neighbours in Rainbow Trout (Oncorhynchus Mykiss) Classification Using Image-Based Features. *Sensors* **2020**, *18*, 1027, doi:https://doi.org/10.3390/s18041027.

[6] Atlam, H.F.; Oluwatimilehin, O. Business Email Compromise Phishing Detection Based on Machine Learning: A Systematic Literature Review. *Electronics* **2022**, *12*, 42, doi:https://doi.org/10.3390/electronics12010042.

[7] Omari, K. Comparative Study of Machine Learning Algorithms for Phishing Website Detection. *International Journal of Advanced Computer Science and Applications* **2023**, *14*, doi:https://doi.org/10.14569/ijacsa.2023.0140945.

[8] Jamal, S.; Wimmer, H.; Sarker, I.H. An Improved Transformer-Based Model for Detecting Phishing, Spam and Ham Emails: A Large Language Model Approach. *Security and privacy* **2024**, *7*, doi:https://doi.org/10.1002/spy2.402.

[9] Mohsin, M.I.; Harun, N.H. Classifying Phishing Websites Using Multilayer Perceptron. *Emerging Advances in Integrated Technology* **2024**, *5*, doi:https://doi.org/10.30880/emait.2024.05.01.008.

[10] Chen, J.; Wang, X.; Lei, F. Data-Driven Multinomial Random Forest: A New Random Forest Variant with Strong Consistency. *Journal of big data* **2024**, *11*, doi:https://doi.org/10.1186/s40537-023-00874-6.

[11] Sarasjati, W.; Rustad, S.; Purwanto; Santoso, H.A.; Setiadi, M. Comparative Study of Classification Algorithms for Website Phishing Detection on Multiple Datasets. *2022 International Seminar on Application for Technology of Information and Communication (iSemantic)* **2022**, 448–452, doi:https://doi.org/10.1109/iSemantic55962.2022.9920475.

[12] Jäger, S.; Allhorn, A.; Bießmann, F. A Benchmark for Data Imputation Methods. *Frontiers in Big Data* **2021**, *4*, doi:https://doi.org/10.3389/fdata.2021.693674.

[13] Galli, S. Feature-Engine: A Python Package for Feature Engineering for Machine Learning. *Journal of Open Source Software* **2021**, *6*, 3642, doi:https://doi.org/10.21105/joss.03642.

[14] Phinzi, K.; Abriha, D.; Szabó, S. Classification Efficacy Using K-Fold Cross-Validation and Bootstrapping Resampling Techniques on the Example of Mapping Complex Gully Systems. *Remote Sensing* **2021**, *13*, 2980, doi:https://doi.org/10.3390/rs13152980.

[15] Miao, J.; Zhu, W. Precision–Recall Curve (PRC) Classification Trees. *Evolutionary Intelligence* **2021**, doi:https://doi.org/10.1007/s12065-021-00565-2.

[16] Singh, P.; Hasija, T.; Ramkumar, K. Machine Learning Algorithms for Phishing Detection: A Comparative Analysis of SVM, Random Forest, and CatBoost Models. *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)* **2024**, 1421–1426, doi:https://doi.org/10.1109/icoici62503.2024.10696365.

[17] Liu, H.; Lang, B. Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. *Applied Sciences* **2021**, *9*, 4396, doi:https://doi.org/10.3390/app9204396.

[18] Hooker, G.; Mentch, L. Bridging Breiman's Brook: From Algorithmic Modeling to Statistical Learning. *Observational Studies* **2021**, *7*, 107–125, doi:https://doi.org/10.1353/obs.2021.0027.

[19] Siahaan, M.I.; Rosmaini, E. Use of Classification and Regression Tree (CART) Method for Classification of Labor Force Participation Levels in Medan City in 2019. *FARABI: Jurnal Matematika dan Pendidikan Matematika* **2022**, *5*, 95–103, doi:https://doi.org/10.47662/farabi.v5i2.386.

[20] Mustafa Abdullah, D.; Mohsin Abdulazeez, A. Machine Learning Applications Based on SVM Classification a Review. *Qubahan Academic Journal* **2021**, *1*, 81–90, doi:https://doi.org/10.48161/qaj.v1n2a50.

[21] Murphy, K.P. *Probabilistic Machine Learning : An Introduction*; The Mit Press: Cambridge, Massachusetts, 2022; ISBN 9780262046824.

[22] Garcia, C.M.; Abilio, R.; Alessandro Lameiras Koerich; de, A.; Barddal, J.P. Concept Drift Adaptation in Text Stream Mining Settings: A Systematic Review. *ACM Transactions on Intelligent Systems and Technology* **2024**, doi:https://doi.org/10.1145/3704922.

[23] Do, N.Q.; Selamat, A.; Krejcar, O.; Herrera-Viedma, E.; Fujita, H. Deep Learning for Phishing Detection: Taxonomy, Current Challenges and Future Directions. *IEEE Access* **2022**, *10*, 1–1, doi:https://doi.org/10.1109/access.2022.3151903.

[24] Diallo, R.; Edalo, C.; Awe, O.O. Machine Learning Evaluation of Imbalanced Health Data: A Comparative Analysis of Balanced Accuracy, MCC, and F1 Score. *STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health* **2024**, 283–312, doi:https://doi.org/10.1007/978-3-031-72215-8_12.