2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI) | 979-8-3315-4066-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICOICI62503.2024.10696365

Machine Learning Algorithms for Phishing Detection: A Comparative Analysis of SVM, Random Forest, and CatBoost Models

Preet Singh
Chitkara University Institute of Engineering and Technology
Chitkara University
Punjab, India
s.preet@chitkara.edu.in

Taniya Hasija
Chitkara University Institute of Engineering and Technology
Chitkara University
Punjab, India
taniya@chitkara.edu.in

KR Ramkumar
Chitkara University Institute of Engineering and Technology
Chitkara University
Punjab, India
k.ramkumar@chitkara.edu.in

Abstract— **Phishing attempts are increasing nowadays due to the advanced internet penetration worldwide. The objective of this research is to create a phishing detection system that is both efficient and effective. This work has taken greater significance in enhancing cybersecurity against the increasing threat of phishing attacks. The proposed approach is for data preparation and applying SMOTE for dataset balancing, training/ testing on a dataset of 11,430 URLs with 87 features extracted from URL structure, content, and external services. The findings indicate that CatBoost gives the best accuracy, thus performing better with fewer misclassifications than Random Forest, evidenced by its confusion matrix. The research concludes that integration herewith of machine learning models more so CatBoost results in phishing detection systems has improved its accuracy and reliability to an accuracy of 97.24%. Such advancement provides a robust solution for mitigating phishing risks.**

Keywords— Cyber Security Infrastructure, Cyber Fraud Prevention, Sustainable Cyber-Security Solutions, Cyber Threat Mitigation.

I. Introduction

Now a common cybersecurity problem, phishing attempts seriously endanger people and companies. Phishers use several dishonest strategies to tempt vulnerable individuals into revealing private information including financial data and login passwords. Usually using weaknesses in website security, these assaults target consumers via emails, instant chats, or hostile websites [1]. The development of efficient detection and prevention strategies depends on a thorough awareness of the nature and features of phishing websites [2]. Usually meant to appear somewhat similar to real sites, phishing websites make it difficult for people to tell real from false pages. Early discovery of phishing websites is essential to minimize any damage and guard consumers against becoming victims of these frauds [3]. Because they can examine complicated patterns and precisely categorize websites as real or fake [4], implementing strong machine learning algorithms such as Support Vector Machines (SVM), Random Forest, and CatBoost in phishing detection systems has proved to be rather useful. These systems provide a wide range of tools and methods that one may use to improve detection accuracy and efficiency. While RandomForest gives insights into feature relevance, CatBoost offers quick training and prediction capability and SVM is well-known for its resilience in handling high-dimensional data [5, 6]. The work intends to create a thorough and effective phishing detection system by combining several algorithms. Apart from spotting phishing websites, incorporating machine learning methods in cybersecurity is essential for protecting consumers' digital experiences. The need for advanced phishing detection systems cannot be emphasized as technology develops [7-9]. This research analyses the use of SVM, RandomForest, and CatBoost in phishing detection and gives a thorough analysis of their performance. The results of this work show how well these algorithms identify phishing websites and provide insightful analysis for further work in this field [10].

The article's organisation follows the order in which Section II provides information about the ML Supervised-Classification approaches: Support Vector Machine (SVM and different kernels), Random Forest and CatBoost. Section III focuses on the study of pertinent literature for the applications of ML in phishing website categorisation. Section IV explains the methodology and approach of the article. This encompasses the dataset's relevant details, the employed classification methodology, the model's assessment via performance evaluation criteria based on the confusion matrix, and the recommended model design for the systematic phishing detection categorisation exploration method. The findings and comparison are presented in Section V. The latter portion of the paper presents the conclusions.

II. Machine Learning

ML is characterized into two main types: Supervised learning and Unsupervised learning. The next section elucidates the supervised classification-based learning technique employed in this study [6].

A. Support Vector Machine (SVM) algorithm

Support Vector Machine is the ML-supervised algorithm utilised for classification and regression tasks. It finds the ideal hyperplane separating data into several classes with the maximum margin to be shown in the high dimensional plane. SVM seeks to separate between phishing and legitimate websites according to acquired characteristics in the framework of phishing detection. SVM is used as an efficient and resilient algorithm to plot the feature in high-dimensional environments [11]. SVM may, however, be computationally demanding, particularly for big datasets, and its performance is much influenced by the kernel and hyperparameter choice.

B. *Random Forest:*

Random Forest is an ensemble learning technique that improves classification accuracy by incorporating multiple decision trees. Every tree learns on a random selection of data and characteristics; the average of all the decision-making results and a majority vote for classification determine the ultimate forecast. Although Random Forest can accommodate large datasets comprising numerous features in phishing detection and is less prone to overfitting than a single decision tree. Its strength is in its capacity to faithfully show complex interactions and linkages among features [7].

C. *Categorical Boosting Algorithm:*

Designed to effectively handle categorical data, CatBoost is a gradient boosting tool. It creates sequential decision trees whereby every new tree seeks to fix the mistakes of the past. CatBoost is notable for its capacity to manage categorical data without much preprocessing required. CatBoost may use several elements of web pages and URLs in phishing detection to raise categorization performance. Although it is well-known for its great accuracy and resilience, like other gradient boosting techniques it can be computationally costly and may need careful hyperparameter adjustment to provide the best results [6].

III. Literature Review

The various relevant literature studied for the classification of Phishing websites is given in Table I. There are some clear limits in the literature study on phishing detection employing ML techniques. The dependence on antiquated or very tiny datasets is one major problem as it could not fairly depict the present scene of phishing attempts. Phishing strategies change constantly; hence datasets must be updated often to keep the detection techniques effective.

Much research also concentrates on a small number of ML algorithms, often excluding a thorough comparison across several methods and settings. This limited emphasis might result in inadequate evaluations of the dependability and applicability of several techniques. Furthermore, restricting us is the use of conventional feature selection techniques, which might not fully exploit hybrid or advanced methodologies for phishing detection.

TABLE I. Relevant Literature Study Summary

Ref No. / Year	Authors	Techniques/Algorithm Employed for Classification	Summary of the Research Paper
[12] / 2020	Rashid et al.	Support Vector Machine (SVM)	This paper proposed a machine learning-based phishing detection technique. SVM achieved 95.66% accuracy in distinguishing phishing and legitimate websites. The technique utilizes features extracted from URLs and HTML source codes. The dataset includes both internal and external characteristics.
[13] / 2006	Wu et al.	Security toolbars, browser address bar, status bar	The study evaluated the effectiveness of security toolbars and other browser elements in preventing phishing attacks. Experiments showed that all tested security toolbars were ineffective, with users being spoofed 34% of the time. The study concluded that poor website design contributed to user deception.
[14] / 2018	Mahajan et al.	Decision Tree, Random Forest (RF), SVM	The research focused on detecting phishing URLs using machine learning. The dataset consisted of 36,711 URLs. Random Forest achieved the highest accuracy of 97.14%. The study emphasized the importance of feature extraction and aimed to identify the most effective ML algorithm for phishing detection.
[15]/ 2020	Jain and Gupta	Logistic Regression (LR), RF	The paper introduced a machine learning-based approach to detect phishing attacks by analyzing hyperlinks. The LR classifier achieved 96.42% accuracy. Feature selection was highlighted as a key component for improving prediction accuracy. The dataset included 10,000 phishing and legitimate web pages.
[16] / 2019	Almseidin et al.	RF	This study used the Random Forest classifier to detect phishing websites, achieving an accuracy of 97.11%. The dataset contained 10,000 phishing and legitimate web pages with 48 features, of which 20 were selected through feature selection. The work highlighted the importance of feature selection.
[17] / 2021	Gandotra and Gupta	RF	The study compared the performance of several ML approaches for phishing detection using a dataset with nearly 5,000 phishing and 6,000 legitimate web pages. Random Forest achieved the best accuracy, with or without feature selection. The study emphasized the efficiency and accuracy of ML-based detection models.
[18] / 2021	Abutaha et al.	SVM, RF, Gradient Boosting Classifier (GBC), Neural Network	The study proposed a technique using lexical analysis of URLs, implemented as a web browser plug-in for phishing detection. The dataset included over a million phishing and legitimate URLs. SVM achieved the highest accuracy of 96.89%, outperforming RF, GBC, and neural network classifiers.
[19] / 2024	Jayaraj et al.	Artificial Neural Networks (ANN), Hybrid Ensemble Feature Selection (HEFS)	This paper presents a machine-learning technique using artificial neural networks for phishing detection. The Hybrid Ensemble Feature Selection (HEFS) method includes a Cumulative Distribution Function gradient (CDF-g) algorithm and data perturbation ensemble. The study focuses on phishing URLs and concludes with a proposed CDG-g-based solution.
[20] / 2011	Zhang et al.	Text-based phishing detection, Neural Networks, Feature Selection	The study developed a text-based phishing detection technique using keyword extraction and Google search. It also proposed a neural network-based classification for detecting malicious web pages. The authors evaluated common feature selection techniques (CFS, wrapper process) and feature space searching techniques (genetic algorithm, greedy forward selection).

IV. Methodology

A. *Dataset*

This work uses 11,430 URLs that were exactly constructed to assess machine learning-based fraud detection systems. Drawn from three primary categories, it comprises 87 distinct features. Examining length, the presence of unique symbols, and subdomain count among other factors, 56 elements derived from the context and semantics of the URLs form the first class. The second class consists of 24 characteristics derived from the content of identical web pages, including analytical HTML elements, keywords, and JavaScript use. The last class consists of seven features—domain age, WHOIS data, web traffic statistics—that were gathered by looking outside services. Most importantly, the dataset is balanced; 50% phishing and 50% genuine URLs are equally distributed, therefore guaranteeing objective model training and evaluation.

B. *Classification Approach*

A subclass of machine learning, classification uses the supervised method to group the labels connected to a dataset into appropriate analytical groups. For instance, grouping related items into separate buckets of classes depending on the characteristics of the groupings and the classes [20] to provide difference. The method applied in this paper for phishing classification is the comparative examination of the SVM classifiers, Random Forest, and CatBoost.

C. *Proposed Architecture*

The research paper's proposed architecture is intended to improve the predictive accuracy of classification models for phishing website categorization by implementing comprehensive procedural steps that commence with data preprocessing and conclude with a comparative analysis of a variety of ML algorithms. First, there is an elaborate data preprocessing phase for the raw dataset. It includes data cleaning for noise and inconsistencies, outlier detection to identify and probably remove anomalous data points, and label encoding to convert categorical variables into a numerical format that machine learning algorithms could process.

After preprocessing, class balancing is done on the dataset using the SMOTE technique to make sure that no class imbalance issues occur and drive the models towards the majority class. The data is then split into training and test sets for the actual phases of training and evaluation. It trains multiple models: an SVM, random forest, and CatBoost—each for their different strengths dealing with different kinds of data.

The models are assessed from post-training using a testing set. The evaluation criteria cover the confusion matrix, accuracy, precision, recall, and F1-score, therefore offering a whole picture of the performance of every model. This thorough assessment makes it possible to compare the performance of several techniques in great detail. Additionally included in the architecture is a comparative analysis stage whereby the most successful model is found by matching the performance measures of the several models. The result of this procedure is identifying a state-of-the-art model that surpasses the others and offers a strong answer to the current issue. Along with great forecast accuracy, this architecture seeks to provide an understanding of the strengths and shortcomings of every model applied.

V. Results and Discussion

A. *Results*

The confusion matrix gives an insight into the correct and incorrect assumptions made by the model for the labels linked to the dataset. In the model evaluation, class-0 implies a legitimate, whereas class-1 indicates a phishing website. Figure 2 gives an insight into the confusion matrix heatmap of SVM. Figure 2 shows that the TP values are 1123, the TN values are 1087, and the model makes 76 erroneous assumptions.

The CatBoost heatmap of the Confusion matrix is determined in fig. 3. The inference drawn from Fig. 3 illustrates the correct classifications made by the model are 1129 for class 0 and 1094 for class 1. The false predictions made by the model are 63. This suggests that the model is good at identifying between phishing and legitimate URLs with minimal errors as high accuracy and balanced precision and recall emerge from this. The minimal false positives and false negatives highlight CatBoost's phishing-detecting ability.

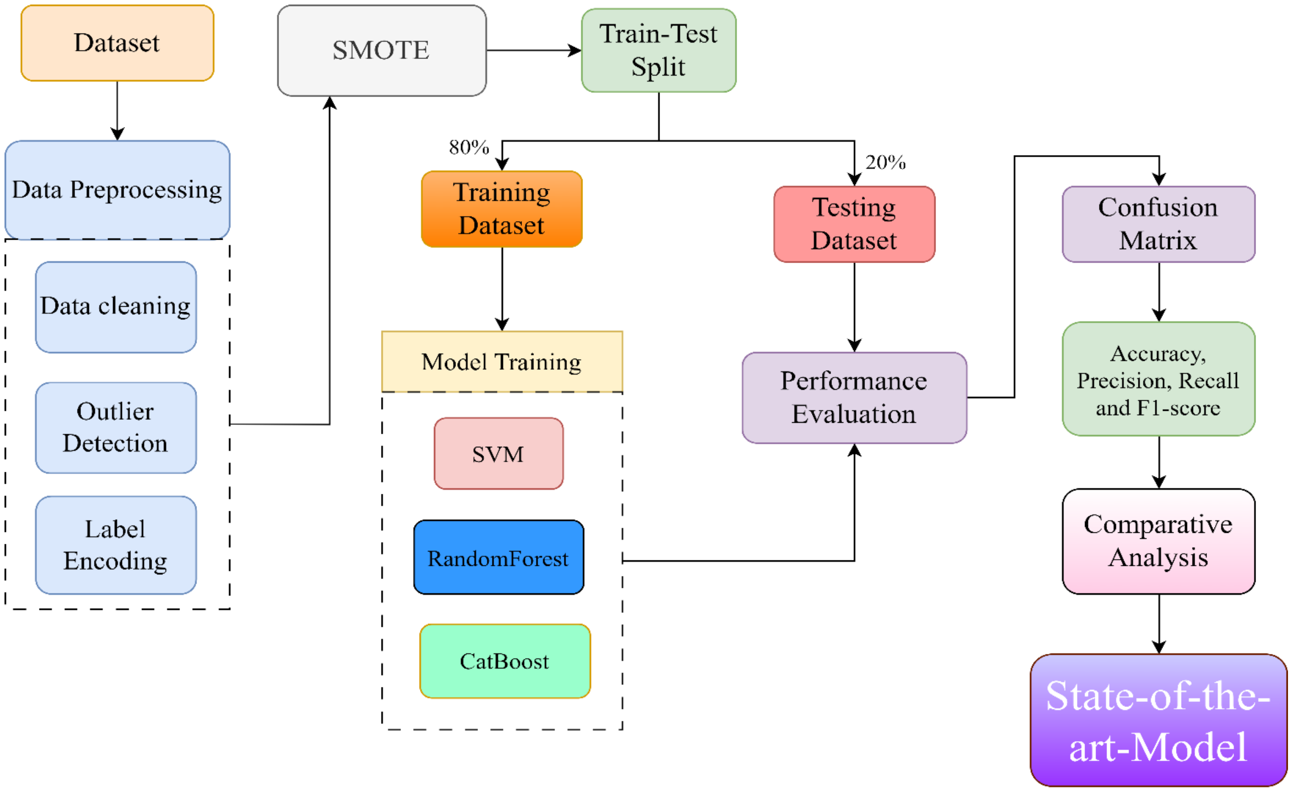

Fig. 1. Architecture Proposed for the Classification of Phishing Sites

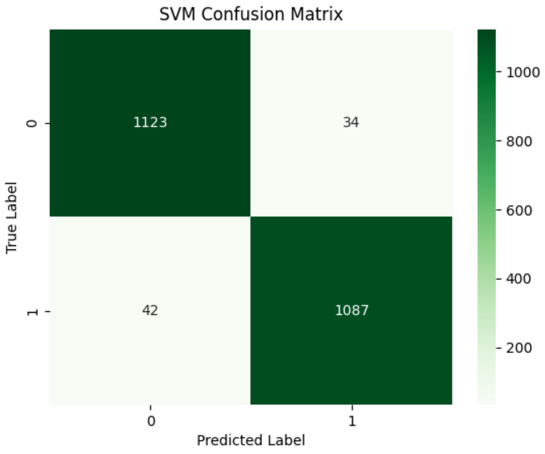

Fig. 2. SVM Confusion Matrix Heatmap

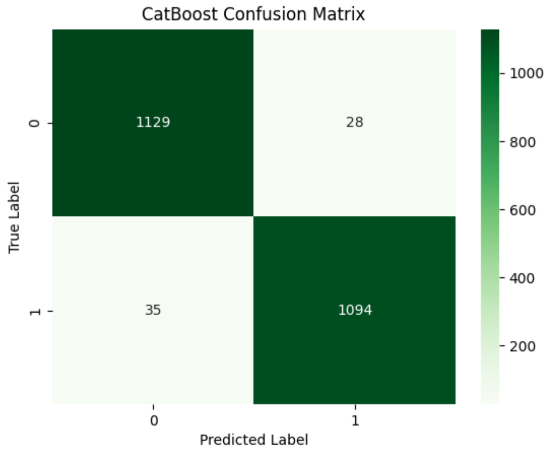

Fig. 3. CatBoost Heatmap Confusion Matrix

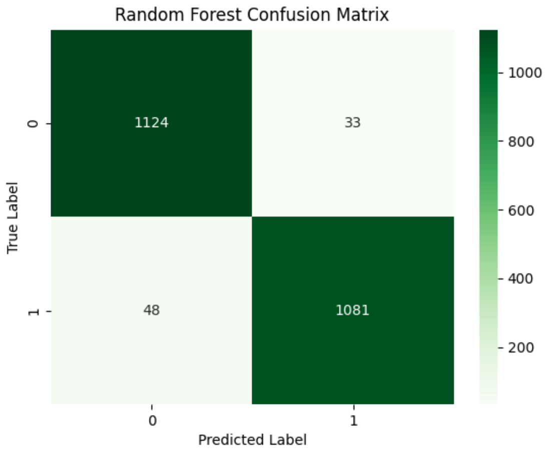

Fig. 4. Random Forest Heatmap Confusion Matrix

Figure 4 provides confusion matrix performance for the Random Forest URLs classification model. It accurately groups 1,041 phishing URLs (true positives) and 1,124 authentic URLs (true negatives). Among the misclassifications are 48 phishing URLs falsely labelled as valid (false negatives) and 33 legal URLs improperly labelled as phishing (false positives). The model shows great accuracy generally even if it shows many more FP and FN than the CatBoost model.

B. Comparative Analysis

The comparative analysis is based on the computation of the various parameters: accuracy, precision, recall and f1-score calculated by the confusion matrix of the respective algorithm. Figure 5 compares the accuracy of the CatBoost, Random Forest and SVM.

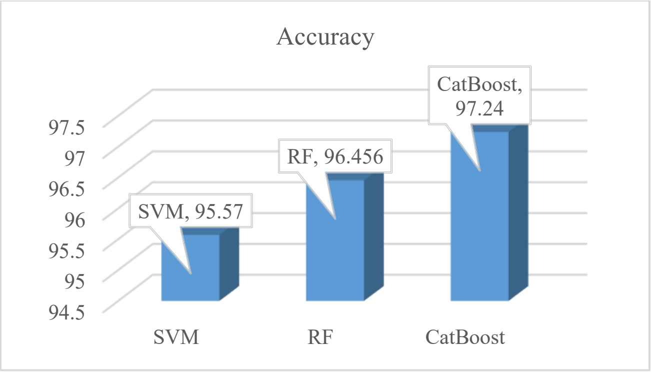

Fig. 5. Comparative Analysis of Accuracy

Figure 5 demonstrates that when comparing the accuracy of the two methods, the CatBoost achieves a higher accuracy of 97.24% compared to the SVM and Random Forest. In Table 1 we can see the class-0 precision, recall, and f1-score values. Classification_report function in sklearn.metric library is used to perform these simulations. For class-0 (legitimate), the CatBoost algorithm outperforms the other models as determined by Fig. 6. Table II determines the model's f1-score, recall, and precision for the category class-1(Phishing) and the comparison plot is given in Fig. 7.

TABLE II. CLASS-0 CLASSIFICATION PERFORMANCE

Model Name	Precision	Recall	F1-Score
SVM	0.9639	0.9706	0.9673
RF	0.9606	0.969	0.9652
CatBoost	0.9699	0.9758	0.9729

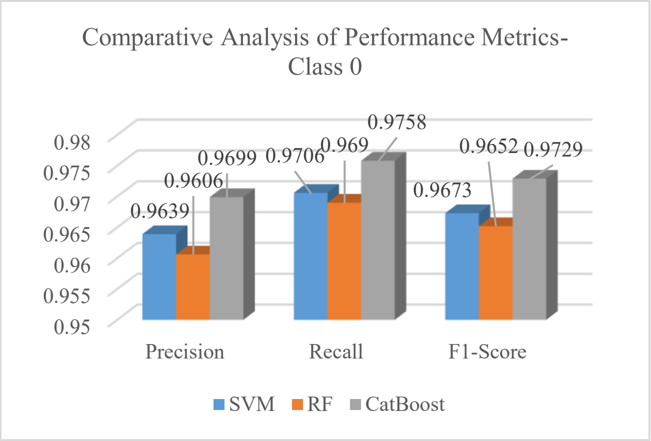

Fig. 6. Comparative Analysis for Class-0

TABLE III. CLASS-1 CLASSIFICATION EFFICIENCY EVALUATION

Model Name	Precision	Recall	F1-Score
SVM	0.9697	0.9628	0.9652
RF	0.9687	0.9593	0.9640
CatBoost	0.9750	0.9690	0.9720

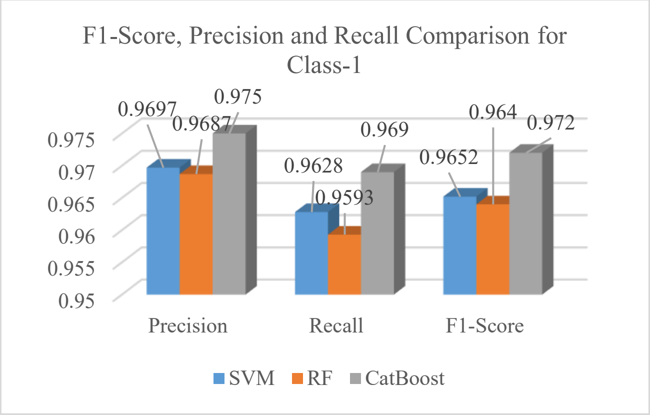

Fig. 7. Comparative Analysis for Class-1

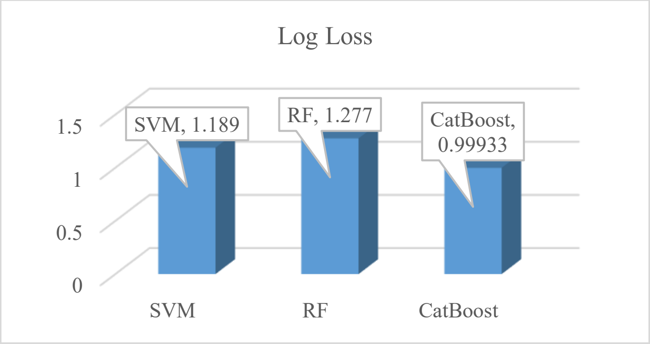

Fig. 8. Comparative Analysis of Log Loss for Class-1

Fig. 8 Log Loss comparison of the algorithm determines that the CatBoost performance is higher and has less loss value. The recall, accuracy, and f1-score evaluation for the distinct classes, class-0 and class-1, indicates that the CatBoost has greater value. Table IV shows the comparison of the proposed model CatBoost algorithm with the existing models. The accuracy of the CatBoost model is higher than the other existing models.

TABLE IV. COMPARISON WITH EXISTING MODELS AND PROPOSED MODEL

Reference	**Accuracy**
[12]	95.66%
[14]	97.14%
[15]	96.42%
[16]	97.11%
Proposed Model (CatBoost)	97.24%

The ROC curve analysis shown in fig. 9 evaluate the ML model's capacity to differentiate between positive and negative classes. The area of the ROC curve's (AUC) performance evaluation was the benchmark. CatBoost obtained optimal class separation from the others with a 1.0 AUC. Both Random Forest and SVM exhibited good performance with AUCs of 0.99; CatBoost came out to be the best model. These findings suggest that for this particular task, CatBoost offers the best accurate classification between positive and negative samples. All three models were rather effective as they excelled in this categorization issue.

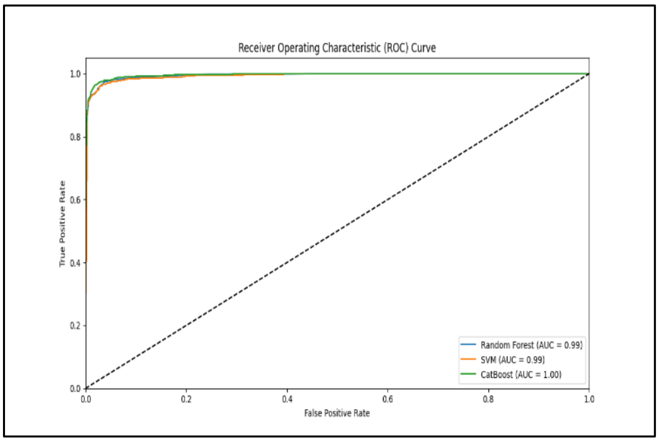

Fig. 9. ROC Comparison of Algorithm

VI. CONCLUSION

This work has proposed an effective phishing detection system by using various machine learning models: Random Forest, Support Vector Machine, and CatBoost. This study solves the phishing challenges that compromise private information and hence judging individuals and organizations as a whole. This effort intends to use machine learning capabilities to increase phishing detection system reliability and accuracy. Using machine learning models, especially CatBoost increases the phishing detection accuracy and reliability, therefore providing a strong and useful method for lowering phishing threats. This study highlighted the significant contribution of machine learning techniques to improve cybersecurity defense against various phishing attempts.

REFERENCES

[1] R. Alazaidah et al., "Website phishing detection using machine learning techniques," Journal of Statistics Applications & Probability, vol. 13, no. 1, pp. 119-129, 2024.

[2] A. A. Akinyelu, "Machine learning and nature inspired based phishing detection: a literature survey," International Journal on Artificial Intelligence Tools, vol. 28, no. 05, p. 1930002, 2019.

[3] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, and S. Hossain, "Phishing attacks detection using machine learning approach," in 2020 third international conference on smart systems and inventive technology (ICSSIT), 2020: IEEE, pp. 1173-1179.

[4] A. Zamir et al., "Phishing web site detection using diverse machine learning algorithms," The Electronic Library, vol. 38, no. 1, pp. 65-80, 2020.

[5] A. K. Jain and B. B. Gupta, "A machine learning based approach for phishing detection using hyperlinks information," J. Ambient Intell. Humanized Comput., vol. 10, pp. 2015-2028, 2019.

[6] S. Mittal, I. Sharma, and A. Kumar, "A Report: Machine Learning and Its Applications," in 2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS), 2023: IEEE, pp. 1-6.

[7] P. Singh, T. Hasija, and K. Ramkumar, "Malware Classification to Strengthening Digital Resilience: Comparing SVM Kernel and Logistic Regression," in 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), 2024: IEEE, pp. 1461-1466.

[8] T. Hasija, V. Kadyan, K. Guleria, A. Alharbi, H. Alyami, and N. Goyal, "Prosodic feature-based discriminatively trained low resource speech recognition system," Sustainability, vol. 14, no. 2, p. 614, 2022.

[9] T. Hasija, K. Ramkumar, A. Kaur, S. Mittal, and B. Singh, "A survey on nist selected third round candidates for post quantum cryptography," in 2022 7th International Conference on Communication and Electronics Systems (ICCES), 2022: IEEE, pp. 737-743.

[10] N. Nagy et al., "Phishing URLs detection using sequential and parallel ML techniques: comparative analysis," Sensors, vol. 23, no. 7, p. 3467, 2023.

[11] A. Kumar, I. Sharma, and S. Mittal, "Enhancing Security through a Machine Learning Approach to Mitigate Man-in-the-Middle Attacks," in 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), 2024: IEEE, pp. 1-6.

[12] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing detection using machine learning technique," in 2020 first international conference of smart systems and emerging technologies (SMARTTECH), 2020: IEEE, pp. 43-46.

[13] M. Wu, R. C. Miller, and G. Little, "Web wallet: preventing phishing attacks by revealing user intentions," in Proceedings of the second symposium on Usable privacy and security, 2006, pp. 102-113.

[14] R. Mahajan and I. Siddavatam, "Phishing website detection using machine learning algorithms," International Journal of Computer Applications, vol. 181, no. 23, pp. 45-47, 2018.

[15] B. B. Gupta and A. K. Jain, "Phishing attack detection using a search engine and heuristics-based technique," Journal of Information Technology Research (JITR), vol. 13, no. 2, pp. 94-109, 2020.

[16] M. Almseidin, A. A. Zuraiq, M. Al-Kasassbeh, and N. Alnidami, "Phishing detection based on machine learning and feature selection methods," 2019.

[17] E. Gandotra and D. Gupta, "An efficient approach for phishing detection using machine learning," Multimedia security: algorithm development, analysis and applications, pp. 239-253, 2021.

[18] M. Abutaha, M. Ababneh, K. Mahmoud, and S. A.-H. Baddar, "URL phishing detection using machine learning techniques based on URLs lexical analysis," in 2021 12th International Conference on Information and Communication Systems (ICICS), 2021: IEEE, pp. 147-152.

[19] R. Jayaraj, A. Pushpalatha, K. Sangeetha, T. Kamaleshwar, S. U. Shree, and D. Damodaran, "Intrusion detection based on phishing detection with machine learning," Measurement: Sensors, vol. 31, p. 101003, 2024.

[20] H. Zhang, G. Liu, T. W. Chow, and W. Liu, "Textual and visual content-based anti-phishing: a Bayesian approach," IEEE transactions on neural networks, vol. 22, no. 10, pp. 1532-1546, 2011.

[21] Kaggle, “Datasets | Kaggle,” *Kaggle.com*, 2019. https://www.kaggle.com/datasets, accessed on 19 July 2024.