This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

A Machine Learning Approach for Phishing Attack Detection

Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil
Terna Engineering College, Nerul, Navi Mumbai, India

Corresponding Author: Rohini Patil, Email: rohiniapatil01@gmail.com

Abstract: Phishing is the easiest method for gathering sensitive information from unwary people. Phishers seek to get private data including passwords, login information, and bank account details. Cyber security experts are actively seeking for trustworthy and effective ways to identify phishing websites. In order to distinguish between legal and phishing URLs, we used machine learning (ML) technology. In this research work using ML technology extraction and analysis of both types of URLs was performed. Extreme Gradient Boosting (XGBoost),Decision Tree (DT), Logistic Regression (LR),Random Forest (RF), and Support Vector Machine (SVM) were used to identify phishing websites. The goal was to identify phishing URLs and determine the most effective ML technique by comparing the accuracy rates of each algorithm. In this, proposed methodology two datasets were used. The accuracy of models was calculated on Phishtank and UCI dataset using kfold, feature selection and hyperparameter tuning method. Performance measures precision, recall and F1-score and Receiver Operating Characteristics (ROC) curve were calculated. RF provided an accuracy of 98.80% and 97.87% on the Phishtank dataset and UCI respectively. Highest precision, recall, F1-score value was 99% each and AUC-ROC value was 99.89% with Phishtank dataset. Validation with other researchers showed better results with proposed methodology. Therefore this methodology can be of help to identify phishing websites.

Keywords: machine learning; phishing; legitimate; malicious; URL; website

I. INTRODUCTION

Phishing is a sort of cyber-attack where attackers pose as legitimate companies or send phoney emails to mislead victims into disclosing personal information like login passwords or financial data. Due to the attackers' tendency to create websites that closely resemble authentic ones and their potential use of social engineering to trick visitors into providing their information, these assaults can be challenging to spot. To combat these attacks, organizations and individuals have developed phishing website detection systems. These systems are designed to automatically detect and flag phishing websites and helping to protect users from falling victim to these attacks. These systems may use a combination of techniques such as analysing structure and content and comparing with known phishing

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

website databases using ML algorithms. Phishing website detection systems have become increasingly important sophisticated in recent years with increasing phishing attacks. These systems can help organizations protect their employees and customers from falling victim to phishing attacks and can also help individuals protect their personal information. With more and more activities moving into the digital space, it is more important than ever to have a good phishing website detection system in place.
To develop a malicious website that closely resembles a legitimate website, phishing has recently become a top worry for security concerns. Getting personal information is the attacker's primary goal. As users are not aware of phishing assaults, the attackers are becoming more successful. It is highly challenging to combat phishing attacks since they prey on user vulnerabilities, yet it is crucial to improve phishing detection methods. Attackers change URLs to appear authentic using encryption and many other straightforward ways in order to escape blacklists. With the help of ML approach an algorithm can examine different blacklisted and valid URLs and their properties in order to precisely identify phishing websites.

The paper is organized as follows. Section 2 describing related work. Section 3 describes methodology used for implementation. The results and discussion are mentioned in section 4, whereas section 5 describes conclusion.

II. RELATED WORK

Pooja and Sridhar [1], introduced a technique for identifying phishing websites by combining Convolutional Neural Network (CNN) with bidirectional Long Short-Term Memory (LSTM) networks. They reported enhancement of the efficiency by combining CNNs and LSTMs by extracting features from website content. System Accuracy with KNN, DT, LR, and XGBoost was 98.40%, 99.05%, 92.08% and 99.8% respectively.

Sindhu et. al [2] demonstrated methods to identify phishing attempts using a variety of ML approaches and developed a system using the RF, SVM, and Neural Network (NN) algorithms in association with back propagation. The system was trained using a dataset of phishing and non-phishing examples and then tested on new examples to evaluate its performance. Using Random Forest, SVM, and neural networks, accuracy of 97.835%, 97.89% and 95.444% respectively was obtained.

Y. Su [3], explained method to identify phishing attempts using a particular ML approach. The authors developed a phishing detection system using LSTM-RNN algorithms. Reported an accuracy of CNN was 97.42% and LSTM was 99.14% respectively.

Nadar et. al [4], described several ML approaches and techniques to create phishing detection systems, and then evaluated their performance using a comprehensive comparison in detecting phishing websites. The paper also outlined various feature extraction techniques used. Challenges and limitations of current methodologies are also described the authors.

Mandadi et. al [5], described the use of several ML algorithms like DT and RF to create a phishing detection system and evaluate its performance using a dataset

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

from Pishtank website. The authors also described the feature extraction methods used.

A comparison of the results obtained from different techniques suggested RF to be the best method for website phishing detection with an accuracy of 87.0%.

Alam et. al [6], described the use of ML approach. The authors also described various feature extraction methods such as REF, Relief-F, IG, and GR Algorithm. They evaluated the performance of DT and RF to detect phishing websites. In their study, RF provided the highest accuracy of 96.96%.

Saha et. al [7], described deep learning techniques.Author discussed feature extraction techniques as REF, Relief-F, IG, and GR Algorithm. This research created deep learning-based phishing detection systems that can accurately recognize and report phishing websites.

Patil et. al [8], described a method for detecting and preventing phishing websites using LR, DT and RF. This system trained the model by feeding it with visual features, heuristic features and blacklist and whitelist approach. Highest accuracy of 96.58% was seen with RF.

Huang et. al [9], described a method for detecting phishing URLs using a combination of CNN and attention-based hierarchical recurrent neural networks (RNN). The system used LR, RF, SVM, CNN & RNN models. Highest accuracy of 97.90% was achieved using combination CNN and RNN.

Vilas et. al [10], described an approach for detecting websites using ML. The method involved training a ML model to classify websites as phishing or legitimate based on their features, such as the URL structure, the content of the website, and the presence of certain keywords using ML models.

Chapla et. al [11], described a method for detecting web phishing using ML and fuzzy logic. The method involved analysing the features of a URL, such as the domain name and the structure of the URL, and using fuzzy logic to determine the likelihood that the URL is a phishing site. Accuracy rate of 91.46% was seen in this study.

Aburrous et. al [12], presented a system using fuzzy techniques. The system used URL-based and content-based features to extract information from phishing websites, and then applied fuzzy logic for analysis and classification purpose.

Yang et. al [13], presented deep learning based system. The system used a combination of visual, structural, and behavioural features to extract information from websites, and then applied algorithms, namely CNN and LSTM to classify the websites as phoney or legit.

Kumar et. al [14], outlined a ML-based approach for identifying phishing websites. The system was trained using Amazon.com verification website. Results showed accuracy rates as follows; KNN: 97.99, DT: 98.02, LR: 97.7%, RF: 98.03% NB: 97.18%.

Singhal et al [15] offered a technique for idea drift detection combined with ML to identify fraudulent websites. The system used URL-based and content-based features to classify the websites. System also monitored the features of the website over time and detects in case of a change or

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

"concept drift" in the feature distribution. Classification algorithms, such as RF, NN and Gradient Boosting (GB) were used to classify the website.

In this study, highest accuracy of 96.4% was observed with the GB model.

A study by Pandiyan S et.al [16] reported accuracy 85% with Light GBM. Using UCI dataset, Alnemari & Alshammari [17] compared accuracy of four models for preventing phishing attacks. In their research, RF model showed accuracy of 96.86% and 97.3%, without and with normalization respectively.

Using three datasets, Mughaid A. et.al [18] demonstrated a very high accuracy with boosted decision tree suggesting its usefulness in detecting phishing attacks. Awasthi & Goel [19] described a phishing detection model using various classifiers and ensemble model. The analysis was performed by combining UCI dataset and Kaggle dataset. These models were tested with and without crossvalidations. Highest accuracy was seen with Extra Trees Classifier.

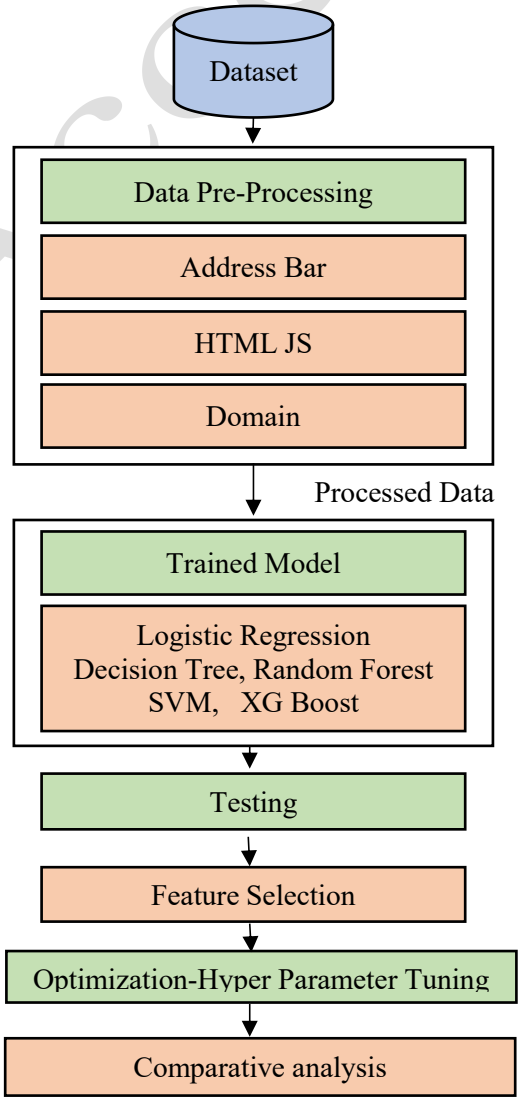
Mohamed et. al [20] used three approaches; ML approach, heuristic-based approach and blacklist-based approach. Among these, ML showed the best performance. Dutta A. [21] described a phishing detection model using RNN and LSTM. The highest accuracy of 97.4% was observed for Phishtank datasets and 96.8% for Crawled dataset, demonstrating better results than the deep learning methods.

III. METHODOLOGY

The methodology is divided into two parts. Part A describes design section, whereas part B describes implementation section.

A. DESIGN

A proposed methodology for a phishing website detection system can vary depending on the specific system being developed. The suggested methodology and flow is depicted in figure 1.



This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

Fig. 1. Proposed System Flow

B. IMPLEMENTATION

The proposed methodology worked on two datasets. The Proposed System is divided into three phases.

1] In the phase 1 we collected a dataset of both trustworthy websites and well-known phishing websites for the system's training and testing purposes.

Dataset Description
Dataset 1 (Phishtank) consisted 10,000 URLs, evenly split between categories, i.e. 5000 phishing URLs collected from Phishtank [22] (https://phishtank.org/ developer_info.php) and 5000 legitimate URLs from the University of New Brunswick(https://www.unb.ca/cic/ datasets /url-2016.html). The phishing URLs were designated with a label of "1". Conversely, all legitimate URLs were labelled as "0". From the datasets, features were extracted using the Python programming language. A total of 25 features were recovered from the jumbled data set, and the model was trained using this set of data. Table 1 shows extracted features and its data type.

Table 1. Feature Description

Sr. No.	Feature	Data Type
1	Domain of URL	Character
2	Having IP Address	bool
3	Has "@" Symbol	bool
4	URL Length	bool
5	URL Depth	int
6	Has embed domain	bool
7	Contains "http/https" in Domain	bool
8	Short Web address Services	bool
9	Has Prefix or Suffix in web address	bool
10	Records in DNS server	bool
11	Network Congestion	bool
12	Domain Age	Bool
13	Expiry of Domain	Bool
14	Redirecting IFrames	Bool
15	Status Bar Customization	bool
16	Blocked Mouse Click	bool
17	Forwarding To Another Web page	bool
18	google_index	bool
19	Count (%)	int
20	Count (?)	int
21	Count (-)	int
22	Count (=)	int
23	Count (.)	int
24	Count (www)	int
25	Label	bool

For dataset 2-UCI, originated from the UCI Machine Learning repository and it composed of 11,055 URLs, 6157 of which were phishing examples and 4898 were legitimate. The URL of UCI [23] is https://archive.ics.uci.edu/ml/datasets/phishing+websites. The outcome was categorized as either 1 (not a phishing attempt) or -1 (a phished URL). Every feature of the URLs was depicted as a column with a value of 1 if the URL was fully phished, 0 if it was partially phished, or -1 if was benign. UCI included a total of 30 features. Additional features other than from Phishtank were Favicon, Using Non-Standard Port, URL of Anchor, Links in <Meta>, <Script> and <Link> tags, Abnormal URL.
The webpages in the dataset were then used to extract characteristics, such as URL patterns, website content, and other characteristics. These features were used to train and test the system.

2] Phase 2 included model training and cross validation. In Model training, the system

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

was then trained using a ML algorithm, such as a LR, DT, RF, XGBoost and SVM on the extracted features and labelled data. The model's performance was assessed using K-fold cross validation, which divided the data into 10 folds.

3] Phase 3 included feature selection and hyperparameter tuning using grid search. Recursive feature elimination was used to select the most relevant features in the data. This step was important to reduce overfitting and improve the model's performance. The model's performance was then optimised through hyperparameter tuning, which involved fine-tuning the model's parameters. Finally, the selected model was evaluated by using various measures. Classifiers used are mentioned below.

SVM: SVM was used for phishing website detection by training a model to classify websites as phishing or legitimate based on various features such as the website's URL, the website's structure, and the website's content. A dataset of tagged legal and phishing websites was used to train the SVM algorithm in this instance. On the basis of the characteristics learnt during training, the model may then classify new websites as authentic or phishing.

RF: RF can be used for phishing website detection. The algorithm was based on DTs, a type of model that can be used to make predictions by following a series of decisions or "if-then" rules. Multiple DTs were trained on a dataset with labelled genuine and phishing websites in the event of an RF. The final prediction was then created by combining the predictions from each tree.

DT: DT Algorithm created a model of decisions and their possible consequences, represented in the form of a branching structure. A DT model may be trained on a dataset of labelled phishing and legit websites in the case of detecting phishing websites.

LR: LR Algorithm can be used for phishing website detection. By applying a logistic function to the data, the method simulated the correlation between variables. For phishing website detection, a LR model was trained on a dataset of labelled phishing and legitimate websites.

XGBoost: XGBoost is a powerful ML algorithm widely used for classification and regression. It is an implementation of gradient boosting framework with DTs as base learners. The main idea behind XGBoost was to iteratively add new DTs to the model for improving its performance. Each new tree was added to correct the errors made by earlier trees in the model. This process is referred to as boosting.

IV. RESULT AND DISCUSSIONS

As per methodology discussed in phase 1 and phase 2, the comparison of Training, Testing and K-Fold Accuracy of both dataset is shown in Table 2.

Table 2. Comparison of Models

Model	Phishtank			UCI		
	Training	Testing	K-fold	Training	Testing	K-fold
SVM	96.45	96.5	96.25	95.38	95.25	94.51
XGBoost	99.13	98.6	98.13	98.6	97	**96.91**
RF	99.65	**98.75**	**98.31**	98.93	**97.65**	96.89

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

DT	**99.68**	97.85	97.84	**99.05**	97.06
LR	93.86	93.85	93.74	92.88	92.72

As mentioned in table 2, the highest training accuracy on Phishtank was observed in DT. Highest testing and k fold accuracy on Phishtank was observed with RF. The lowest K-fold accuracy was seen by the LR model of 93.74%. The graphical representation of each classifier is shown in figure 2.

Fig. 2. Comparison of Models

As feature selection is a vital step. Considering this, feature selection using Recursive Feature Elimination (RFECV) method was used. Hyperparameter tuning helps to optimise the proposed methodology. Comparison of K-fold, Feature selection and Hyperparameter tuning using Grid Search CV on Phishtank are mentioned in Table 3.

Table 3. Comparison of K-Fold, Feature Selection and Hyperparameter Tuning

Model	K-fold	Feature Selection	Hyper parameter Tuning
SVM	96.25	96.50	98.30
XGBoost	98.13	98.55	98.55
RF	**98.31**	**98.75**	**98.80**
DT	97.84	97.95	97.95
LR	93.74	94.00	94.00

Highest K-fold, feature selection and Hyperparameter Tuning accuracy of 98.31%, 98.75% and 98.80% respectively was observed in RF model. The lowest feature selection accuracy was seen with the LR model at 94%. The graphical representation of each model is shown in figure 3.

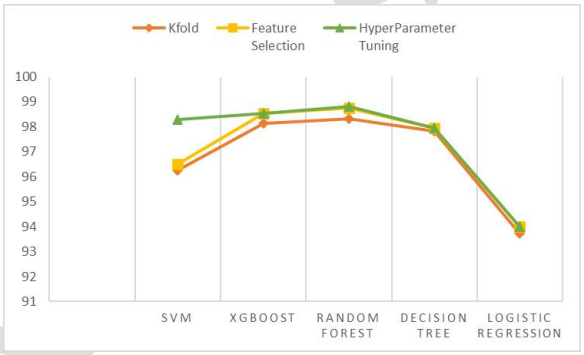


Fig. 3. K-Fold, Feature Selection and Hyperparameter Tuning Accuracy

Table 4 shows comparison of K-fold, Feature selection and Hyperparameter tuning accuracy on UCI. The highest K-fold accuracy of 96.1% was seen with XG boost model while maximum feature selection and hyperparameter tuning accuracy were shown by RF classifier. The lowest accuracy was shown in the LR classifier.

Table 4. UCI - Comparison of K-Fold, Feature Selection and Hyperparameter Tuning accuracy

Model	K-fold	Feature Selection	Hyper parameter Tuning
SVM	94.51	95.07	95.98
XGBoost	**96.91**	97.00	97.30
RF	96.89	**97.65**	**97.87**
DT	96.19	96.83	96.83

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

LR	92.70	92.76	92.76

Figure 4 and figure 5 represents ROC curve analysis of both Phishtank and UCI.

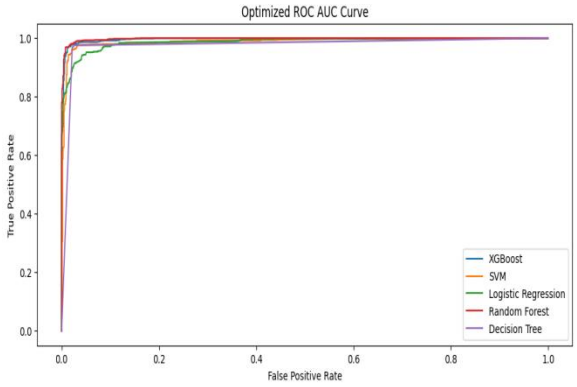


Fig. 4. Phishtank - ROC AUC Curve

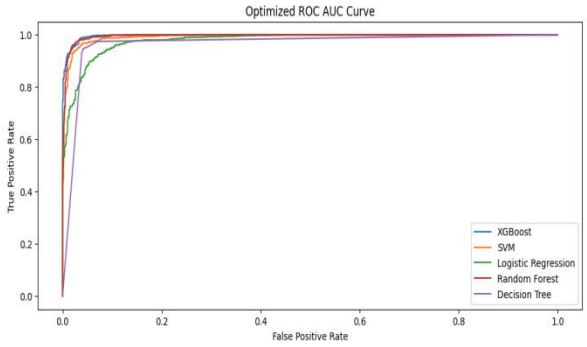


Fig. 5. UCI - ROC AUC Curve

Performance analysis on various measures observed by the best model on Phishtank and UCI dataset is shown in figure 6.

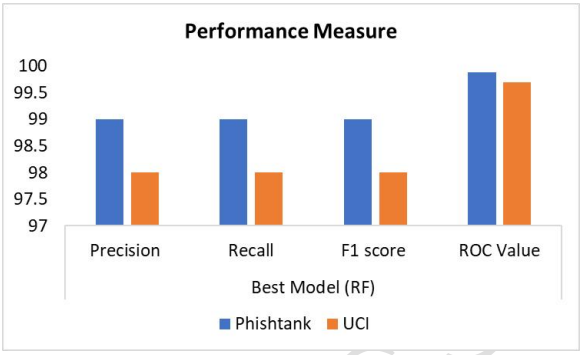


Fig. 6. Comparative performance analysis- Phishtank and UCI dataset

Validation with contemporary researcher

The results of our proposed methodology validated with the contemporary researchers who worked on Phishtank are mentioned in Table 5. The proposed methodology yielded good results with an accuracy of 98.80% in comparison with the other researcher.

Table 5. Validation - Phishtank

Ref no.	Model	Accuracy
Mandadi et. al [5]	RF	86
Alam et. Al [6]	RF	97
Huang et. Al [9]	CNN+RNN	97.9
Kumar et. al [14]	RF	98.03
Proposed Methodology	**RF**	**98.8**

Proposed approach results tested on UCI dataset and validated with the existing researcher are shown in Figure 7. Methodology adopted by current work of RF classifier and existing research results are similar.

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

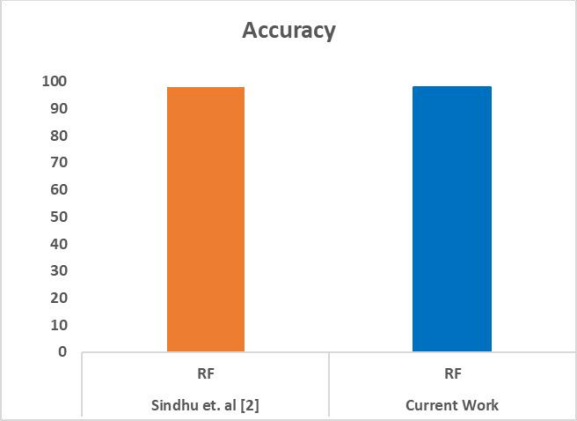


Fig. 7. Accuracy Based Validation

IV. CONCLUSION

It is important for a phishing website detection system to have a high accuracy rate of detecting phishing attempts. The objective of this research was to detect phishing attacks by analysing patterns using machine learning. We observed that the use of feature selection and hyperparameter tuning can help to improve the model accuracy. Proposed ML based framework outperformed in terms of accuracy precision, recall F1-score and ROC value on both datasets. Validation of ML algorithms outperformed other state-of-the-art methods yielding high results in terms of accuracy. Performance measures in proposed framework helped in detecting phishing attacks.

References

[1] A. S. S. V. L. Pooja and M. Sridhar, "Analysis of Phishing Website Detection Using CNN and Bidirectional LSTM," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2020, pp. 1620-1629, Doi: 10.1109/ICECA49313.2020.9297395.

[2] S. Sindhu, S. P. Patil, A. Sreevalsan, F. Rahman and M. S. A. N., "Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation," 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Bengaluru, India, 2020, pp. 391-394, doi: 10.1109/ICSTCEE49637.2020.9277256.

[3] Y. Su, "Research on Website Phishing Detection Based on LSTM RNN," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2020, pp. 284-288, doi: 10.1109/ITNEC48623.2020.9084799.

[4] V. K. Nadar, B. Patel, V. Devmane and U. Bhave, "Detection of Phishing Websites Using Machine Learning Approach," 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-8, doi: 10.1109/GCAT52182.2021.9587682.

[5] A. Mandadi, S. Boppana, V. Ravella and R. Kavitha, "Phishing Website Detection Using Machine Learning," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-4, doi: 10.1109/I2CT54291.2022.9824801.

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

[6] M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. -E. -. Ulfath and S. Hossain, "Phishing Attacks Detection using Machine Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1173-1179, doi: 10.1109/ICSSIT48917.2020.9214225.

[7] I. Saha, D. Sarma, R. J. Chakma, M. N. Alam, A. Sultana and S. Hossain, "Phishing Attacks Detection using Deep Learning Approach," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 1180-1185, doi: 10.1109/ICSSIT48917.2020.9214132.

[8] V. Patil, P. Thakkar, C. Shah, T. Bhat and S. P. Godse, "Detection and Prevention of Phishing Websites Using Machine Learning Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697412.

[9] Y. Huang, Q. Yang, J. Qin and W. Wen, "Phishing URL Detection via CNN and Attention-Based Hierarchical RNN," 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 2019, pp. 112-119, doi: 10.1109/TrustCom/BigDataSE.2019.00024.

[10] M. M. Vilas, K. P. Ghansham, S. P. Jaypralash and P. Shila, "Detection of Phishing Website Using Machine Learning Approach," 2019 4th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT), Mysuru, India, 2019, pp. 384-389, doi: 10.1109/ICEECCOT46775.2019.9114695.

[11] H. Chapla, R. Kotak and M. Joiser, "A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier," 2019 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2019, pp. 383-388, doi: 10.1109/ICCES45898.2019.9002145.

[12] M. Aburrous, M. A. Hossain, F. Thabatah and K. Dahal, "Intelligent Phishing Website Detection System using Fuzzy Techniques," 2008 3rd International Conference on Information and Communication Technologies: From Theory to Applications, Damascus, Syria, 2008, pp. 1-6, doi: 10.1109/ICTTA.2008.4530019.

[13] P. Yang, G. Zhao and P. Zeng, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," in IEEE Access, vol. 7, pp. 15196-15209, 2019, doi: 10.1109/ACCESS.2019.2892066.

[14] J. Kumar, A. Santhanavijayan, B. Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104161.

This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.
Citation information: Tarun Choudhary, Siddhesh Mhapankar, Rohit Bhddha, Ashish Kharuk, and Rohini Patil, A Machine Learning Approach for Phishing Attack Detection, *Journal of Artificial Intelligence and Technology* (2023), DOI: https://doi.org/10.37965/jait.2023.0197

[15] S. Singhal, U. Chawla and R. Shorey, "Machine Learning & Concept Drift based Approach for Malicious Website Detection," 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, 2020, pp. 582-585, doi: 10.1109/COMSNETS48256.2020.9027485.

[16] Pandiyan SS, Selvaraj P., Burugari V., Benadit P, Kanmani P."Phishing attack detection using Machine Learning", Measurement:Sensors,2022, Vol.24, https://doi.org/10.1016/j.measen.2022.100476.

[17] Alnemari S, Alshammari M. Detecting Phishing Domains Using Machine Learning. App.Sciences.2023; 13(8):4649. https://doi.org/10.3390/app13084649

[18] Mughaid, A., AlZu'bi, S., Hnaif, A. et al. An intelligent cyber security phishing detection system using deep learning techniques. Cluster Comput 25, 3819–3828 (2022). https://doi.org/10.1007/s10586-022-03604-4

[19] Awasthi, A., Goel, N. Phishing website prediction using base and ensemble classifier techniques with cross-validation. Cybersecurity 5, 22 (2022). https://doi.org/10.1186/s42400-022-00126-9

[20] Mohamed G, Visumathi J, Mahdal M, Anand J, Elangovan M. An Effective and Secure Mechanism for Phishing Attacks Using a Machine Learning Approach. Processes. 2022; 10(7):1356. https://doi.org/10.3390/pr10071356

[21] Dutta AK (2021) Detecting phishing websites using machine learning technique. PLoS ONE 16(10): e0258361. https://doi.org/10.1371/journal.pone.0258361

[22]PhishTank Homepage https://phishtank.org/developer_info.php accessed Aug 20 (2022)

[23] UCI Homepage https://www.unb.ca/cic/datasets /url-2016.html accessed Oct 10 (2022).