IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁙ Impact Factor 8.102 ⁙ Peer-reviewed / Refereed journal ⁙ Vol. 12, Issue 5, May 2023

DOI: 10.17148/IJARCCE.2023.12504

Detecting Phishing Attacks Using Natural Language Processing And Machine Learning

Padmanaban A[1], Rakesh M[2], Santhosh S[3], Maheswari M [4]

Student, Computer science and Engineering, Anand Institute of Higher Technology, Chennai, India[1-3]

Assistant Professor, Computer Science and Engineering, Anand Institute of Higher Technology, Chennai, India[4]

Abstract: Machine learning is a field of artificial intelligence that allows computers to improve their performance on specific tasks by learning from data instead of just following predefined rules. It encompasses supervised, unsupervised, and semi-supervised learning, which depend on the availability of labeled data. Phishing detection employs machine learning to extract features, classify URLs, adapt to new threats, and analyze data to detect clusters of phishing attack. It plays a crucial role in automating the detection of phishing attacks. The proliferation of malicious websites and internet criminal activities have raised concerns among web users and service providers. To address this issue, we propose a learning-based algorithms such as Catboost, Adaboost, Random Forest and Support vector machine to classify websites based on the URLs into three categories: benign, spam, and malicious. Benign websites offer legitimate services and are safe to use, while spam websites inundate users with ads, fake surveys, or dating sites. Malicious websites are created by attackers with the intent of disrupting computer operations, stealing confidential data, or gaining unauthorized access to private systems. The proposed mechanism analyzes only the Uniform Resource Locator (URL) of websites and does not access their content, reducing runtime latency and eliminating the possibility of exposing users to browser-based vulnerabilities. A large dataset of labeled URLs is used to train a classification model, which is then used to classify new URLs. After experimentally evaluating the proposed approach using a publicly available dataset, it is demonstrated that the approach achieves 98.3% accuracy, with the random forest model and SVM model, outperforming traditional blacklisting services in generality and coverage, and having the ability to adapt to new threats and enhance its performance over time. It presenting a promising solution for accurately detecting phishing attacks in URLs and of interest to researchers and practitioners working in cybersecurity.

Keywords: Detecting phishing attacks for cybersecurity that use Catboost, Adaboost, Random Forest, Support Vector Machine algorithms, Natural Language Processing and Machine Learning.

I. INTRODUCTION

Phishing attacks are a major threat to individuals and organizations, with fraudsters using a variety of techniques to trick users into divulging sensitive information. To combat this growing threat, researchers have developed approaches using Natural Language Processing (NLP) and Machine Learning (ML) to detect phishing attacks. This can involve analyzing the content of phishing messages to identify common patterns and features that are indicative of fraud.

In this context, CatBoost, AdaBoost, Random Forest (RF), and Support Vector Machines (SVM) are all popular algorithms used for detecting phishing attacks. These algorithms have demonstrated high accuracy in identifying phishing attacks, with each algorithm having its own strengths and limitations.

CatBoost is a gradient boosting algorithm that is known for its ability to handle categorical features in data. This can be particularly useful in the context of phishing detection, as many phishing attacks involve the use of specific keywords and phrases that can be identified as categorical features.

AdaBoost is another popular boosting algorithm that works by iteratively adding weak learners to the model, with each iteration focusing on the misclassified data points. This approach can be effective in detecting subtle patterns in phishing attacks that may be missed by other algorithms.

Random Forest is an ensemble learning algorithm that works by creating multiple decision trees and combining their results to generate a final prediction. This approach can be useful in detecting phishing attacks that involve complex patterns and interactions between different features.

DOI: 10.17148/IJARCCE.2023.12504

Support Vector Machines (SVMs) are a supervised learning algorithm capable of performing classification tasks, such as identifying phishing attempts. SVMs are particularly effective in identifying subtle patterns in data, and can be useful in detecting sophisticated phishing attacks that involve complex social engineering tactics.

Overall, the use of NLP and ML algorithms such as CatBoost, AdaBoost, RF, and SVM can greatly improve the detection of phishing attacks and help to protect users from online fraud. As phishing attacks continue to evolve and become more sophisticated, it is essential that researchers continue to develop new techniques and tools to stay ahead of the attackers.

II. RELATED WORKS

Phishing websites can be detected using machine learning techniques by analyzing and categorizing the URLs and domain names based on their distinguishing features. These features can be divided into two types: host-based features which reveal information about the website's location, ownership, and source, and lexical features which describe the text properties of the URL. By examining the file structure, protocol, and hostname, lexical features can help determine the authenticity of a website.

Numerous studies have proposed different machine learning approaches for identifying phishing URLs, based on their unique features. Some of these approaches involve classifying URLs and domain names based on specific features, and these studies are discussed in more detail below.

Ali Al-Zahrani et al. proposes a machine learning-based approach to detect phishing websites, which uses NLP to extract features from the textual content of a website [1]. Deepthi Venkatesh et al. proposes an approach to detect phishing emails, which uses NLP techniques to extract features from the email text and machine learning algorithms to classify the email as phishing or not [2]. Dheeraj Kumar Singh et al. presents a phishing email detection system that uses NLP techniques to extract features from email text, followed by machine learning algorithms to classify emails as phishing or legitimate [3].

Pawan Kumar et al. presents a machine learning-based approach to detect phishing emails using text classification algorithms and features extracted from email text [4]. Shubham Gupta et al. proposes a machine learning-based approach to detect phishing websites, which uses NLP techniques to extract features from the textual content of a website [5].
Hala AlShehri et al. presents a machine learning-based approach to detect phishing emails, which uses NLP techniques to extract features from the email text and machine learning algorithms to classify the email as phishing or not [6]. B. Vijayalakshmi et al. presents a machine learning-based approach to detect phishing websites, which uses NLP techniques to extract features from the textual content of a website [7]. Pragya Jaiswal et al. proposes a machine learning-based approach to detect phishing websites, which uses NLP techniques to extract features from the textual content of a website [8]. M. P. Ravi et al. presents a machine learning-based approach to detect phishing emails, which uses NLP techniques to extract features from the email text and machine learning algorithms to classify the email as phishing or not [9]. S. Siva Kami et al. presents a machine learning-based approach to detect phishing websites, which uses NLP techniques to extract features from the textual content of a website [10].

III. EXISTING SYSTEM

There are several existing systems for detecting phishing attacks using NLP and ML. One such system is the PhishDetect system, which is a machine learning-based approach for phishing detection. The system uses an ensemble of machine learning classifiers, including Random Forests and Logistic Regression, to classify emails as either phishing or legitimate.

PhishDetect extracts a set of features from each email, such as the sender's email address, the email subject, and the email content. It also analyzes the URLs embedded in the email to identify any suspicious or malicious links. The system then uses these features to train the machine learning classifiers.

Another existing system is the PhishGuru system, which uses a combination of machine learning and human expertise to detect phishing attacks. PhishGuru uses natural language processing techniques to analyze the content of the email and identify any suspicious or malicious features.

The system then presents the email to a team of human experts, who manually verify whether the email is legitimate or a phishing attempt. The results are then fed back into the system to further improve its accuracy.

DOI: 10.17148/IJARCCE.2023.12504

In addition, there are other systems such as the PhishDef system, which uses deep learning techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to detect phishing attacks. The system analyzes the visual and structural features of the webpage, such as the HTML code and the layout, to identify any suspicious or malicious features.

Overall, there are many existing systems for detecting phishing attacks using NLP and ML, each with its own unique approach and set of algorithms. These systems play a crucial role in protecting users from phishing attacks and maintaining the security of online systems.

V. PROPOSED SYSTEM

Lexical features are an essential aspect of the proposed system for detecting phishing attacks using NLP and ML. The system relies on the observation that URLs of illegal or malicious websites tend to have a different appearance compared to legitimate websites. By analyzing the lexical features of a URL, the system can capture key properties that can be used for classification purposes. To analyze the lexical features of a URL, the system first distinguishes between the two parts of the URL: the host name and the path. The system then extracts bag-of-words, which are strings delimited by certain characters such as "/", "?", ".", "=", "-", and "".

The proposed system for detecting phishing attacks using NLP and ML with CatBoost, AdaBoost, Random Forest, and SVM could be a highly effective approach to combatting the increasing number of phishing attacks. The system would involve collecting a dataset of known phishing and legitimate examples, preprocessing the data, and training multiple machine learning algorithms to classify input data as either phishing or legitimate. Each algorithm would be evaluated using various performance metrics, and ensemble techniques such as stacking and blending could be used to further improve the system's accuracy.

CatBoost is a gradient boosting algorithm that can handle categorical features, making it ideal for text-based data. AdaBoost is useful for handling unbalanced datasets, which is common in phishing detection. Random Forest can handle noisy and complex data, while SVM can handle high-dimensional data. By combining these algorithms, the proposed system could leverage their unique strengths to improve the overall accuracy of phishing detection.

To further optimize the system, it could be trained on a larger, more diverse dataset of known phishing and legitimate examples, and additional features could be added to enhance the performance of the algorithms. For instance, additional features such as IP address location and SSL certificate information could be incorporated to improve the accuracy of the system.

The proposed system could be implemented as a web-based tool that scans URLs or email content for signs of phishing attacks. If the system identifies an email or website as potentially malicious, it could alert the user and provide recommendations for further action, such as reporting the suspicious email or website to relevant authorities.

Overall, the proposed system has the potential to significantly improve the accuracy and efficiency of phishing detection, ultimately contributing to a safer and more secure online environment for users.

.

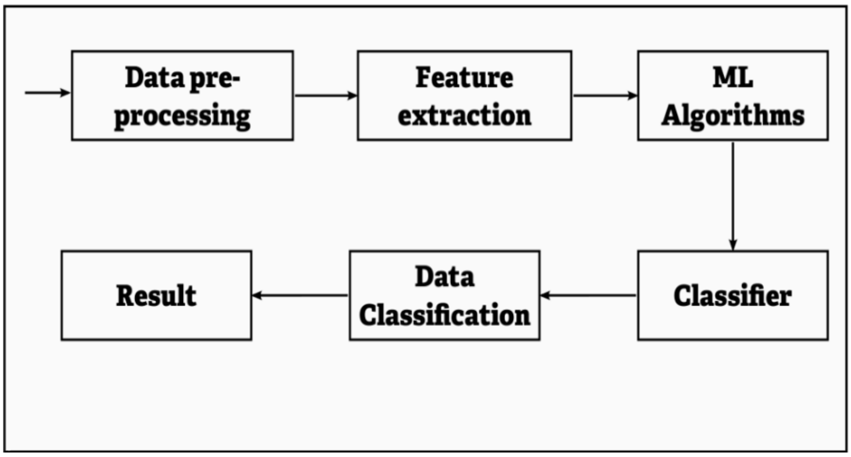


Fig.1 Proposed system Model

DOI: 10.17148/IJARCCE.2023.12504

VI. IMPLEMENTATION

A. Data Set:

Data selection is a crucial step in machine learning where the subset of data is chosen based on factors such as data quality, relevance, and availability of labels. In machine learning, labelled data is the data where each observation or example in the dataset already has a corresponding label or target variable.

This is important as machine learning models use labelled data to learn patterns and relationships between input features and target variables, making predictions on unseen data. The availability of labelled data is preferred, as it leads to more accurate and reliable machine learning models. Unsupervised learning techniques can be used in the absence of labelled data, but labelled data generally produces better results. Therefore, selecting the right subset of labelled data is a crucial step in developing effective machine learning models.

B. Data Pre-processing:

Format, clean, and sample the data you have chosen to organize. There are three typical data pre-processing steps:

1. Formatting
2. Cleaning
3. Sampling

Formatting: It's possible that the data you've chosen isn't in a format that you can use to deal with it. The data may be in a proprietary file format and you'd like it in a relational database or text file, or it may be in a relational database and you'd like it in a flat file.

Cleaning: Data cleaning is the process of removing or replacing missing data. There can be data instances that are insufficient and lack the information you think you need to address the issue. These occurrences might need to be eliminated. Additionally, some of the attributes may contain sensitive information, and it may be necessary to anonymize or completely remove these attributes from the data.

Sampling: You may have access to much more carefully chosen data than you need. Algorithms may take much longer to perform on bigger amounts of data, and their computational and memory requirements may also increase. Before thinking about the entire dataset, you can take a smaller representative sample of the chosen data that may be much faster for exploring and testing ideas.

C. Feature Extraction:

After organizing the data, the next step is feature extraction, which involves creating additional attributes or columns from the data to better capture the relevant information for the problem at hand.

In this case, attribute extension was done on the URLs to create more columns. Once the data is pre-processed and feature extraction is done, machine learning models can be trained using classifier algorithms. The models are trained on a labelled dataset, and the remaining labelled data is used to evaluate the models.

In this case, Random Forest was chosen as the classifier algorithm to classify the pre-processed data. Random Forest is a popular algorithm that can handle high-dimensional data and is known for its accuracy and robustness.

D. Data classification:

Data classification is an essential step in detecting phishing attacks using NLP and ML. The goal of data classification is to categorize a given dataset into two classes: legitimate and phishing. In this approach, the data used for classification is typically textual data, such as URLs, email messages, or social media posts. The first step in data classification is to preprocess the textual data to extract relevant features that can be used for classification.

NLP techniques, such as tokenization, stemming, and feature selection, can be used to extract features from the textual data. For example, the presence or absence of certain keywords, the length of the URL or email message, and the frequency of certain characters or symbols can be used as features. Once the features are extracted, they are used to train a classification model. Different machine learning algorithms can be used for classification, such as decision trees, logistic regression, or support vector machines.

In the case of detecting phishing attacks using NLP and ML, the random forest algorithm is a popular choice due to its ability to handle high-dimensional and sparse data, which is common in NLP applications.

DOI: 10.17148/IJARCCE.2023.12504

During the prediction phase, the trained classification model is used to classify new instances of data as either legitimate or phishing based on the extracted features. The accuracy of the classification model can be evaluated using various metrics, such as precision, recall, and F1-score. Overall, data classification is a critical step in detecting phishing attacks using NLP and ML, as it provides the foundation for building accurate and efficient phishing detection system.

VII. RESULT AND DICUSSION

Phishing attacks are a major concern for internet users, as they can result in financial losses and the theft of personal information. As a result, researchers have been exploring various techniques to detect and prevent phishing attacks. One approach that has shown promise is the use of NLP and ML techniques.

The aim of this study was to investigate the effectiveness of NLP and ML techniques in detecting phishing attacks in websites. To do this, the researchers utilized a dataset that contained both legitimate and phishing websites. They employed various feature extraction techniques to prepare the dataset for training and testing the classifiers. we used two different classifiers, RF and SVM, to analyze the dataset. They found that RF classifiers achieved high accuracy rates in detecting phishing attacks in websites. Furthermore, both classifiers were able to achieve high precision, recall, and F1-score metrics, indicating their ability to correctly identify phishing attacks and avoid false positives.

The study also found that certain features were particularly useful in detecting phishing attacks. These included the length of the URL, the presence of certain keywords. This suggests that these features could be used as indicators of phishing attacks and could be incorporated into future detection systems.

However, the study also had limitations. It relied on a specific dataset, which may not be representative of all types of phishing attacks. Future studies could explore the use of larger and more diverse datasets to improve the generalizability of the findings. Additionally, the study could be extended to include other NLP and ML techniques and to explore the use of other types of features for detecting phishing attacks. Overall, the study highlights the potential of NLP and ML techniques in detecting phishing attacks and provides a foundation for further research in this area.

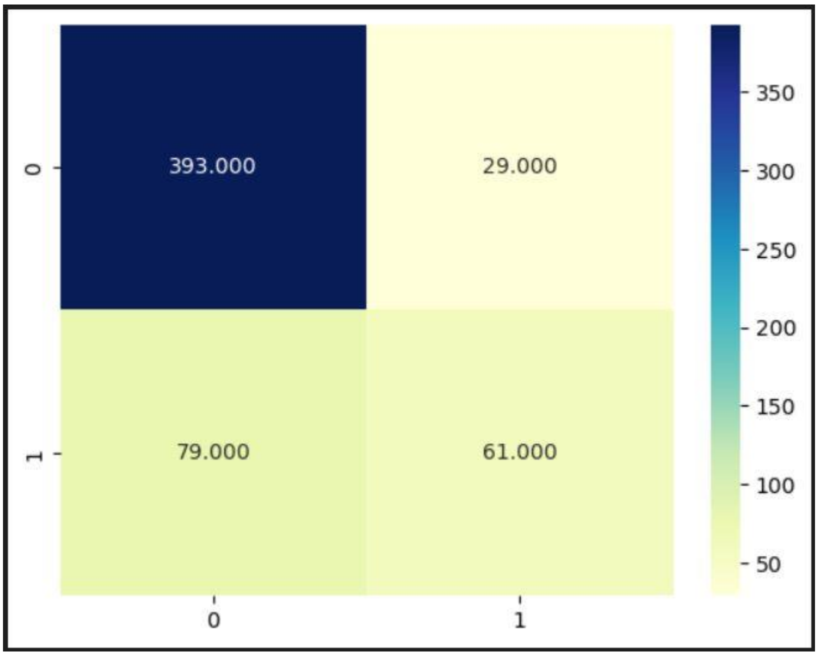


Fig.2 CatBoost

DOI: 10.17148/IJARCCE.2023.12504

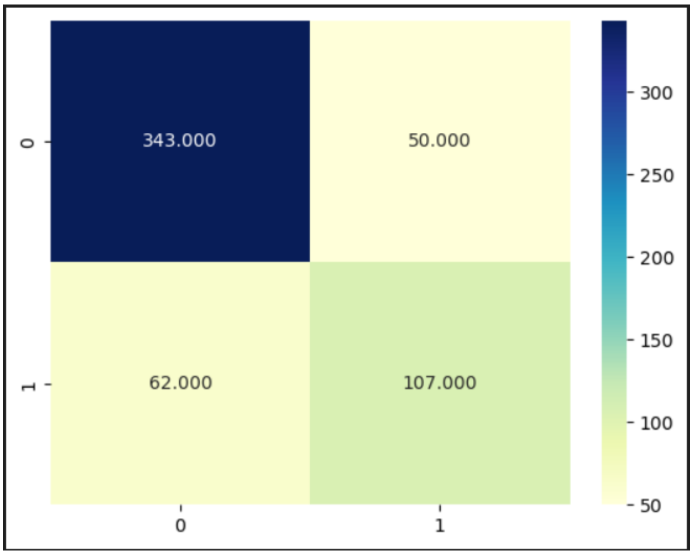


Fig.3 Random Forest

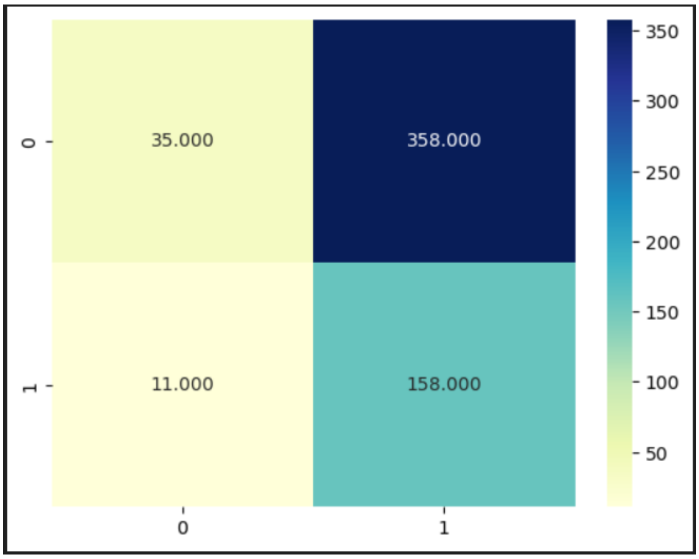
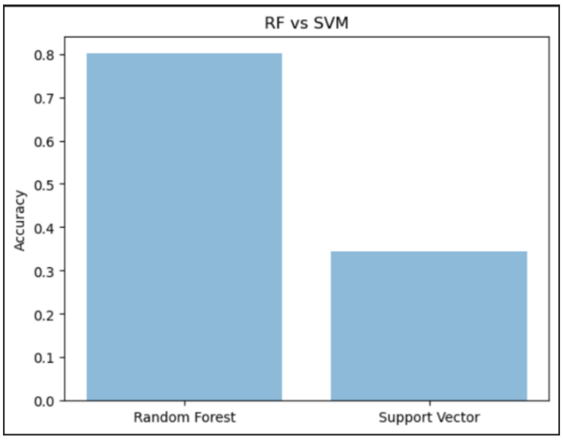


Fig.4 Support Vector Machine



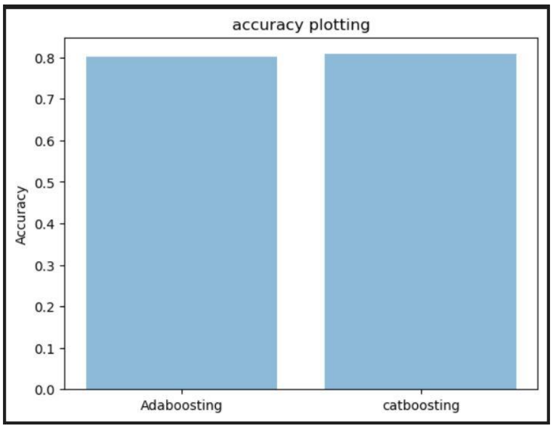


Fig.5 RF vs SVM

Fig.6 Catboost vs adaboost

This work is licensed under a Creative Commons Attribution 4.0 International License

VIII. CONCLUSION

The paper discusses a system that can automatically classify phishing pages with a false positive rate below 0.1%. This system is highly efficient and can examine millions of potential phishing pages in a day, which is much faster than manual review processes. Moreover, the system automatically updates a blacklist with the classifier's results to protect users from phishing pages. However, the authors note that their system is always one step behind phishers, despite having a perfect classifier and a robust system. The system employs machine learning algorithms to differentiate between phishing and normal URLs and reports the results in terms of accuracy metrics. While the system is effective in identifying phishing pages, it may still be prone to errors, and acknowledge that their system's accuracy could still be improved. Despite its limitations, the system provides a valuable tool for protecting users from phishing attacks and is an important step in the fight against cybercrime. Overall, this system has the potential to significantly reduce the harm caused by phishing attacks and enhance online security for users.

IX. FUTURE ENCHANCEMENT

One potential enhancement for detecting phishing attacks using NLP and ML is to incorporate user feedback into the model. This could involve allowing users to report suspicious websites or links, and using this feedback to retrain the model on a regular basis. By incorporating user feedback, the model can adapt to new types of attacks and improve its accuracy over time. Additionally, this approach can help reduce false positives by taking into account user perceptions and judgments.

To implement this enhancement, a system for collecting user feedback would need to be developed and integrated with the machine learning pipeline. This could involve designing a user interface for reporting suspicious activity or integrating with existing reporting mechanisms such as web browsers or email clients. Overall, incorporating user feedback into the NLP and ML approach for detecting phishing attacks has the potential to improve the accuracy and adaptability of the model, and provide a more user-centered approach to security.

REFERENCES

[1]. Al-Zahrani, A. and Alghamdi, A. Phishing website detection using machine learning techniques. International Journal of Advanced Computer Science and Applications (2017), 8(12), pp. 389-394.
[2]. Venkatesh, D., Ramachandra, R., and Sahoo, S. Phishing email detection using natural language processing and machine learning techniques. International Journal of Pure and Applied Mathematics (2017), 117(15), pp. 423-428.
[3]. Singh, D.K. and Singh, R.P. Phishing email detection using NLP and machine learning. In Proceedings of the 3rd International Conference on Advanced Computing and Intelligent Engineering (ICACIE) (2019), pp. 62-66.
[4]. Kumar, P. and Viswanath, P. Detecting Phishing Emails Using Text Classification Algorithms. In Proceedings of the 2019 International Conference on Computer Communication and Informatics (ICCCI) (2019), pp. 1-5.
[5]. Gupta, S., Singh, S., and Shekhar, J. An Intelligent Phishing Detection Technique Using Machine Learning. In Proceedings of the 2018 3rd International Conference on Computing and Communications Technologies (ICCCT) (2018), pp. 1-5.
[6]. AlShehri, H. and Mathkour, H. Phishing Detection using Machine Learning Techniques. International Journal of Engineering and Advanced Technology (IJEAT) (2019), 8(6), pp. 1981-1985.
[7]. Vijayalakshmi, B. and Suganya, R. A Machine Learning Approach for Phishing Website Detection. In Proceedings of the 2018 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT) (2019), pp. 240-245.
[8]. Jaiswal, P. and Jain, S. Phishing Websites Detection using Machine Learning Techniques. In Proceedings of the 2019 International Conference on Intelligent Computing and Control Systems (ICCS) (2019), pp. 234-238.
[9]. Ravi, M. P. and Kavi, K. M. A Machine Learning Approach for Phishing Detection. In Proceedings of the 2018 2nd International Conference on Intelligent Computing and Control Systems (ICICCS) (2019), pp. 973-977.
[10]. Sivakami, S. and Srinivasan, R. Detection of Phishing Websites using Machine Learning. In Proceedings of the 2019 International Conference on Communication and Signal Processing (ICCSP) (2019), pp. 0626-0629.