LITERATURE SURVEY

The Literature Survey is used to provide a brief overview and explanation about the reference papers. Literature survey conveys the technical details related to the project in a proper and detailed manner. A literature survey is a piece of discursive prose, not a list describing or summarizing one piece of literature after another. It is an iterative process, assessing and distilling information.In this paper, the authors proposed a system with a collection or set of Hybrid features to classify websites based on machine learning algorithms. The main feature set is extracted using the cumulative distribution gradient technique, while the data perturbation ensemble technique is used to extract the secondary feature set.

The proposed system detected phishing websites using a machine learning algorithm The Machine learning methods set included eight methods are logistic algorithm,support vector machine,k near neighbor,decision tree, Ada boost, Random forest XG boost, Artificial neural network based on the website structure and was choosen after a comparative study by authors. we get very good performance in ensembling classifiers namely, Random Forest, XGBoost both on computation duration and accuracy.Besides, we will explore to propose and develop a new mechanism to extract new features from the website to keep up with new techniques in phishing attacks.[1]

The authors made a relative study to detect phishing website URLs with machine learning and deep learning algorithms. Random forest , decision tree, Support vector machine formed the architecture of the classification model. Scikit-learn tool has been used to import Machine learning algorithms.This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate.[2]

In this paper, the authors designed a browser extension to detect phishing websites Phishing website is one of the internet security problems that target the human vulnerabilities rather than software vulnerabilities. It can be described as the process of attracting online users to obtain their sensitive information such as usernames and passwords.We have selected the Random Forest technique due to its good performance in classification. As a result, we conclude our paper with accuracy of 98.8% and combination of 26 features.[3]

Authors made a comparative study of various machine learning algorithms Such as (RNN-GRU) Recurrent neural network - Gated recurrent units, Random forest ,Logistic regression to implement an efficient phishing website detection system. In the ML prediction module the author compared multiple machine learning modules using several datasets from the expremental results the RNN-GRU module uptime of the highest accuracy of 99.`18% demonstrating the feasibility of the proposed solutiuon.[4]

This paper proposes a phishing website detection method using reduces feature classification. The extracted features were analyzed using the decision tree, Naïve Bayesian classifier, support vector machine (SVM), and neural network . The classifiers were tested with a data set containing 1,353 real world URLs where each could be categorized as a legitimate site, suspicious site, or phishing site. The results of the experiments show that the classifiers were successful in distinguishing real websites from fake ones over 90% of the time.[5]

The author proposed a phishing website detection method using reduces feature classification. The phishing website has evolved as a major cybersecurity threat in recent times. In recent times machine learning techniques have been used in the classification and detection of phishing websites. In, this paper we have compared different machine learning techniques for the phishing URL classification task and achieved the highest accuracy of 98% for Naïve Bayes Classifier with a precision=1, recall = .95 and F1-Score= .97. [6]

The authors in this paper proposed a phishing detection with using four kind of classifier such as Random forest, Multi layer preceptron, J48, K nearest neighbor. The performance of proposed system implements the machine learning classifiers and has correctly classified phishing using relevent features.proposed approach high accuracy in classifying the phishing.[7]

The authors made a relative study to detect phishing website URLs with machine learning and deep Reinforcement learning algorithms. The proposed model is capable of adapting to the dynamic behavior of the phishing website and thus learn the features associated with phishing website detection.[8]

The author presented a system for phishing website detection using different machine learning algorithms and a bag of words technique. The dataset was imported and the classifier was trained using the features extracted by a content-based approach. The training classifier included machine learning algorithms like Support Vector Machine (SVM) and Naive Bayes (NB) for training and testing purposes. Out of the two algorithms implemented, Support Vector Machine (SVM) showed higher accuracy achieving 95 percent. [9]

The algorithm used for training the classifier is Random Forest in association with ensemble learner identifies the phishing websites with a precision of 94.6 percent. [10]

The authors made a relative study to detect phishing website URLs with machine learning and deep learning algorithms. Convolution Neural Network (CNN) and CNN Long Short-Term Memory (CNN-LSTM) with Logistic Regression formed the architecture of the classification model. The system was designed using tools like TensorFlow along with Keras for machine learning and deep learning model. The dataset was imported from multiple sources to provide better scalability. The phishing website URL dataset was obtained from OpenPhish and Phishtank, while the malicious or spam website URLs were imported from MalwareDomains. [11]

In this paper, the authors proposed a novel approach using machine learning algorithms to detect phishing websites. The model utilized multiple machine learning algorithms for training and testing purposes. In total 30 features were extracted from the imported dataset obtained from various repositories and third-party service providers. The classifier was trained using Random Forest (RF), Decision Tree (DT), Generalized Linear Model (GLM), Gradient Boosting (GBM), and Generalized Additive Model (GAM). Out of all the machine learning algorithms, the Random forest classifier achieved an accuracy of 98.4 percent in the testing phase. [12]

The authors in this paper propose a model to classify websites as legitimate or phishing. The model was implemented in MATLAB and the Data Set was imported from the UCI Irvine machine learning repository. The system comprises of extraction of features from websites using Extreme Learning Machine (ELM), Naïve Bayes (NB), and Support Vector Machine (SVM). Among the algorithms used, the Extreme Learning Machine (ELM) obtained an accuracy of 95.34%. The model was implemented in MATLAB and the Data Set was imported from the UCI Irvine machine learning repository. [13]

Authors in this paper proposed a system named BaitAlarm for determining phishing websites. The presented model used the visual content-based approach for feature representation and classification purposes. The visual features included HTML, JavaScript, and CSS features extracted from the list of websites obtained from the mported dataset. The system relies on the visual content and layout based features of the website to detect whether the website is phishing or legitimate. [14]

The authors proposed a system to detect phishing using heuristic-based methods and feature extraction. The c4.5 decision tree algorithm was used for analysis and computing the heuristic values to determine whether a website is legitimate or phishing. The Dataset was imported from Phishtank and Google, proposed using the trained classifier, and used for detection purposes. The model achieved an accuracy of 89.40%. [15]

The paper proposed a flexible decision filter to extract and classify features from the inputted website URL. The system was implemented using a neural network model and other optimizers included AdaDelta, Adam, and Stochastic Gradient Descent (SGD). The Dataset was imported from Phishtank and Chainer was used to develop and implement the model. Among the three optimizers used, Adam obtained an accuracy of 94.18%. [16]

The authors in this paper presented a feature selection methodology to detect phishing website detection. The dataset was obtained from the UCI Irvine machine learning repository. Various algorithms were implemented by the authors for training purposes and after a comparative study, the authors finally concluded that different classification methods and strategies showed different results. Based on classification strategies and the data mining techniques the execution outcome results are incremented or decremented. [17]

This paper proposed a system that determines phishing mails using two existing systems, Machine Learning Anti- Phishing System (MLAPT) and Phishzoo. The Phishzoo system uses the visually based approach for phishing detection while the Machine Learning Anti- Phishing System (MLAPT) helps in determining the mails present on the system into a phishing or benign category. The presented model proved effective to manage personal sensitive information on social networking websites. [18]

The paper proposes an efficient way to detect phishing websites using a URL identification strategy utilizing the approach of the Random Forest algorithm. Phishtank was used to gather the required dataset. Out of the total 30 features listed, only 8 features were used for parsing to analyze the feature classification. The system model was partitioned into three stages which consisted of classification, parsing, and analysis. The model achieved 95% accuracy for the Random forest algorithm implemented using Rstudio. [19]

Authors designed a system with a detection technique involving a fresh approach for phishing website detection named PhishLimiter. The proposed system used Deep Packet Inspection (DPI) along with Software-Defined Networking (SDN) through web communications and emails for identifying malicious activities. The real-time DPI and phishing signature classification based on SDN programmability provided PhishLimiter, the flexibility to address phishing attacks in real-time. This also helped in better network traffic management and evaluated attacks in real-world environments proving an effective solution to identify phishing attacks. [20]

In this paper, the authors used artificial intelligence techniques like neural networks for detecting phishing websites. The obtained data set from third-party service providers was divide into two parts, each for a specific purpose. The training module imported 80 percent of the dataset while the remaining 20 percent was used for the Testing phase. The Neural network model utilized the input of 17 neurons to compare with 17 characteristics in the imported dataset. The system determined whether the website is legitimate or phishing based on one hidden layer level of processing and output of two neurons. The proposed system showed an accuracy of 92.48 percent. [21]