

A Semantically Analysis Deep Learning and Feature Selection Technique to Detect Phishing URLs Attack

Preeti¹ and Priti Sharma²

¹⁻²Maharshi Dayanand University Rohtak Haryana, India

Email: Miskhokhar121@gmail.com, priti@mdurohtak.ac.in

Abstract—Phishing is an attack that targets internet users, organizations, e-commerce websites, financial institutions, and government institutes on the internet. To resolve this problem, many techniques are used, like heuristic-based, DL (Deep Learning-Based) and ML (Machine Learning-Based). This paper provides a detailed understanding of all DL models and their feature selection and extraction techniques. Feature extraction techniques play an important role in reducing the dimensions of existing datasets. This study is very helpful to the researchers to analyse which model and feature extraction technique are best to resolve the problem of phishing attacks on the internet.

Index Terms— Phishing Attack, Deep learning, Feature Extraction technique.

I. INTRODUCTION

Phishing is a cyber-attack based on social engineering (Alper Ozcan) [11], which aims to abstract user personal information like account numbers, credit or debit card numbers, Facebook account information, and so on. Typically, the attacker creates the fake website and then sends an email to the target audience instructing the recipient to click on a link on the fake website [11]. These emails contain various kinds of viruses, such as worms, malware, and so on. Nowadays, phishing is extensively used in social media, traditional media, and research literature. Since different researchers give distinct definitions of phishing attacks, according to the APWG [30], "Phishing is a crime employing both social engineering and technical subterfuge to steal consumers' identity data and financial account credentials." Because of varied definitions, "phishing" is a scalable act of deception whereby impersonation is used to obtain information from a target "[31]. Fig. 1 Described "APWG saw 384,291 attacks in March 2022, which was the highest monthly total in APWG's reporting history. In the first quarter of 2022, APWG observed 1,025,968 total phishing attacks. In the area of research, there are many techniques to resolve the problem of phishing attacks, like character-based techniques, content-based techniques, genetic algorithm-based techniques, machine learning-based techniques, and deep learning techniques. All techniques have their own advantages and disadvantages. So we can't say we have a single technique to resolve the problem of phishing attacks. Machine learning is a good technique to solve the problem of phishing attacks, but it has some limitations. In machine learning, you extract the feature manually, etc. Apart from this, deep learning is also good for resolving the problem of phishing attacks. In detail, it is described in Section II. The main goal of this study is to analyze the performance of DL models like RNN, CNN,

LSTM, BI-LSTM, and others. Also, describe various types of feature selection and extraction techniques. Feature selection and extraction play an important role in building the model. In feature selection, consider Variance Threshold, Feature Selection with Correlation, Mutual Information in Classification, and the Chi-square Test for Feature Selection.

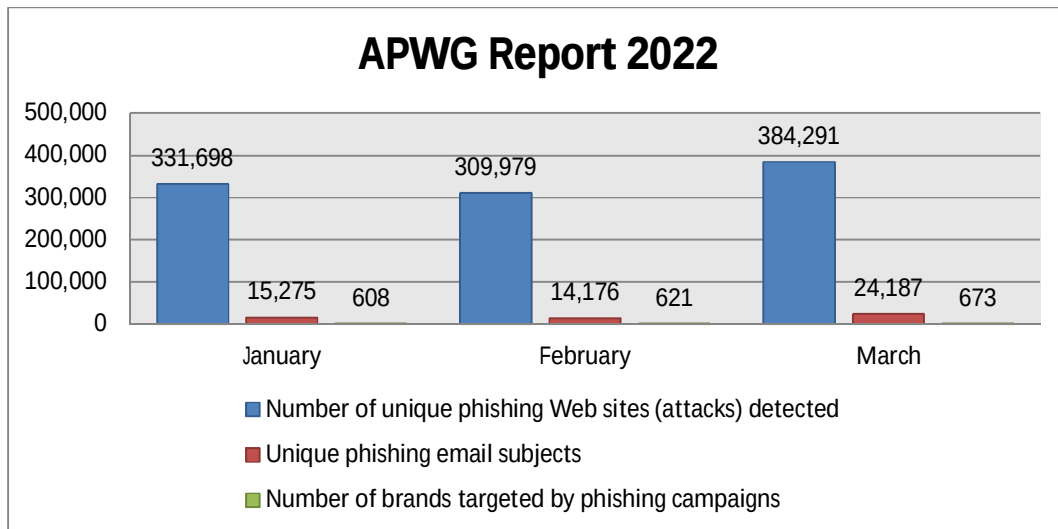


Fig.1. APWG Report of Phishing Attack (6 month 2022)

II. LITERATURE REVIEW

BU Seok-Jun: In this paper, Propose a model using DL(Deep Learning) and genetic algorithms for feature selection. This study is split into three steps. In the first step, select the feature of URLs. In the second step, select the genetic algorithm for feature extraction. After applying the genetic algorithm, select fifteen features from URLs to propose the model. The third step used CNN and LSTM to propose the model. The primary goal of this study is to enhance recall (sensitivity). Recall and accuracy improved by 4.13% and 7.07%, respectively. used a dataset containing 222,541 URLs.

Ali Aljofey : In this study, they proposed a model using CNN (convolution neural network) and the character level embedding technique. This considers the features of the URL rather than website content and any third-party services. They used sequence pattern features for the classification of the URLs. URL feature classification is classified into four groups. These are hand-crafted features; character level count vector features; character level TF-IDF features. This paper collects the datasets from Phish-Tank, Common-crawl, and Alexa. For the evaluation of the proposed model, they used four datasets (D1,D2,D3,D4).The D1 dataset contains 318,642 URLs.

Chidimma Opara: In this study, they proposed the model using CNN. Which classifies the web page as a basic HTML document? Extract the features from the HTML document. used word and character embedding techniques to extract the features. The dataset contains 50,000 HTML documents. The proposed model gives 93% accuracy. In an HTML document, consider the web page text, images, tables, hyperlinks, tables, and lists. HTML-Phish learns the semantic meaning using characters and words present in the HTML document. A Web Crawler is used to collect the HTML document data. To create a parser and extract the features of an HTML document, we used the soup library in Python. **Feature work:** In the future, compare the research work with the feature engineering-based model that is used to extract the feature from the HTML document. Also, try to make extensions and plug-ins for the web browser to implement in a real-time environment.

RASYMAS, Tomas: This paper proposed the model using CNN and LSTM. For feature selection and extraction, we used lexical features, character-level embedding, and word embedding techniques. Which proposed model gives 94.4% accuracy? which uses the ReLU activation function. It contains 60 lexical features of URLs. Collect the data from Phishtank.com.

YANG PENG: This paper proposed a model using multidimensional features with a deep learning model. Multidimensional features contain the features of URLs, Web text features, and webpage code features. The proposed model gives an accuracy of 96.78% and the false positive rate is 0.59%. Collect that dataset from

Phish-tank.com containing 1021758 phishing URLs and 989021 legitimate URLs from Dmoz.net. For the classification, they used the soft-max function for classification. It uses CNN and LSTM models to propose a model for phishing URL detection. In the feature, try to make extensions and plug-ins for this browser to solve problems in a real-time environment.

Hung Le :- This paper proposed the model using CNN. To extract the features of datasets, they use word and character level embedding techniques. Firstly, test the model based on the character-level embedding technique and then apply the word-level embedding technique. On the basis of accuracy, combine word and character-level embedding techniques. Collect the data from

Virus-Total The dataset contains 14,050,275 benign URLs and 949,725 malicious URLs. To train the model, it uses lexical patterns that contain the primary domain, path, and file extension. The proposed model gives an AUC of 0.9929.

In this paper, we proposed the model using the Multifaceted Function Phishing detection method. They developed multidisciplinary apps to detect phishing websites. This app contains URL data samples, website textual information, homepage software features, and classification outcomes based on deep learning techniques. For feature extraction, the proposed model uses the character encoding technique.

Rundong Yang:- This paper proposed the "Multidimensional Phishing Susceptibility Prediction Model" (MPSPM). For the experiment, they selected 1105 volunteers and identified volunteers as susceptible or not susceptible. This used seven supervised learning methods. The proposed model gives an accuracy of 89.04%. The main aim of this paper is to identify and analyze those factors that affect network susceptibility. In this study, they proposed a model using the DNN (Deep Neural Network). For feature extraction, they used the character embedding technique. The proposed model has good accuracy.

In the feature, consider the feature of the web page to propose a model to detect a phishing attack.

Umer Ahmed Butt: In this study, they proposed the model using SVM, NB, and LSTM. The accuracy of the proposed model is SVM: 92.622% NB: 95%, while LSTM is 96%..The Linear Kernel is a good choice for text classification. For dataset-preparation include Data Integration, Data cleaning, Data transformation, Normalization, and Smoothing techniques. A confusion matrix is used to evaluate the performance of a proposed model. To extract the features using regular expression and NLP-Kit in Python programming.

Alper Ozcan :A deep neural network and short-term memory were used in this paper's model, which was proposed. NLP and character embedding were both employed for feature extraction and selection. Two datasets were utilized during model training. 90.79 percent Accuracy is obtained based on the first dataset. 91.79% accuracy is obtained based on the second dataset set. To increase accuracy, the suggested model employed a hybrid approach that combined DNN and BI-LSTM. Limitation: New datasets are only suggested after careful examination of existing datasets. A noisy data set may have been utilized to train the model. The Word embedding technique was utilized to extract the features in order to improve the suggested model's accuracy.

III. RESEARCH GAP

S.No	Author	Technology	Feature selection technique	Accuracy	Limitation	Feature work
1	Seok-Jun Bu[1]	CNN, LSTM	Genetic Algorithm	Improve Accuracy 4.13%p and Recall 7.07%p.	Consider only URL-based features.	In feature consider the text, HTML, And image features to propose the model.
2	Ali Aljofey[2]	CNN	Character embedding technique.	95.02%with D1 dataset and 98.58%, 89.46%, and 85.22% on the Benchmark dataset.	Consider only URL features.	Consider using the multidimensions dataset to increase the suggested model's effectiveness.
3	Chidimman Opara	CNN	Character and word embedding techniques are used.	Accuracy 98% and 2% False Positive Rate.	Not used in a real-time environment.	Make extension add with browser to recognize phishing websites in real-time.

4	Tomas RASYMAS	CNN AND LSTM	Character and word embedding techniques are used.	Accuracy is 94.4%.	Consider only URL features.	In feature use more methods on the same dataset.
5	PENG YANG	CNN AND LSTM	CNN AND LSTM	96.78%.	Not used in a real-time environment.	To use in a real-time environment make an extension and plug-in with a web browser.
6	Hung Le	CNN	Character and word embedding techniques are used.	AUC-0.9929	Used only URL features.	Use Multidimensional features to propose the model.
7	P. Vigneshwaran	Multifaceted function Phishing detection method	Character embedding Technique	90.25%	Take more training time.	To increase the accuracy of features used the word embedding technique.
8	Rundong Yang	Seven Supervised Machine learning model.	Character embedding technique.	88.27%	Analysis the result is based on the outcome of 1105 volunteers.	To enhance the accuracy of features using deep learning techniques.
9	Mohammed Alshehri	DNN	Character embedding technique, word-level encoding,	88.13%	Doesn't contain the feature related to web content.	In the feature, consider the feature of the web content to propose the model.
10	Umer Ahmed Butt	SVM, LSTM, NB	Character level embedding technique is used	SVM-92.62% NB-95% LSTM-96%	Doesn't contain the feature related to web content.	In the feature, consider the feature of the web content to propose the model.
11	Alper Ozcan	LSTM, DNN	Character level embedding, NLP feature extraction.	91.79%	Collected dataset may be noisy.	The feature used the word embedding technique.

IV. DEEP LEARNING TECHNIQUES

DL is a branch of ML. It is often used to simulate the behaviour of the human brain. In deep learning, neurons train based on different datasets and test them. Deep learning is a more powerful technique to figure out the problem of phishing URL detection because it creates neural networks, which are a combination of the input layer, hidden layer (as per problem definition), and output layer. The architecture of deep learning is defined in Fig.1. In this have three layer.

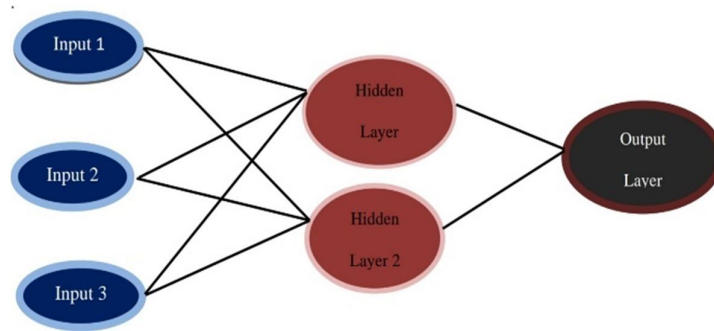


Fig 1. The Architecture of Deep Learning

A. Deep Learning Model Type

Deep learning models use several algorithms. While no one model is perfect to resolve the problem of phishing attacks, each model has some advantages as well as some disadvantages. Fig. 2 shows various types of deep learning models to resolve the problem of phishing attacks.

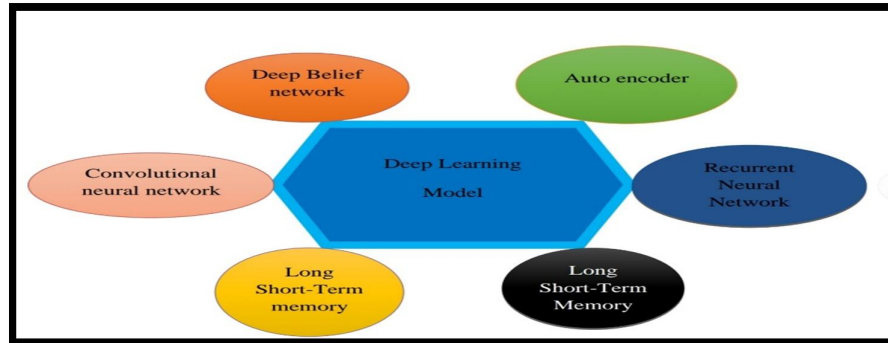


Fig 2. Type of Deep Learning Models

B. Deep Belief- Network

Yi and colleagues [12] It proposes the model using the DBN model of deep learning. To train the model using two features: interactivity and original features. This proposed model gives a 0.6% false positive rate and a 90% true positive rate.

Maneesh Kumar Verma [13] used DBN and ANN to propose a model to detect phishing attacks. The proposed model employs an input layer, multiple hidden layers, and output layers. These layers are connected to each other. There is no connection between the units within each layer. To train the model using supervised and unsupervised methods [14]. In a supervised way, DBN works with respect to the classifier and performs classification. DBN can learn from the knowledge and extract the features in the best way. Fig 3. Show the architecture of DBN.

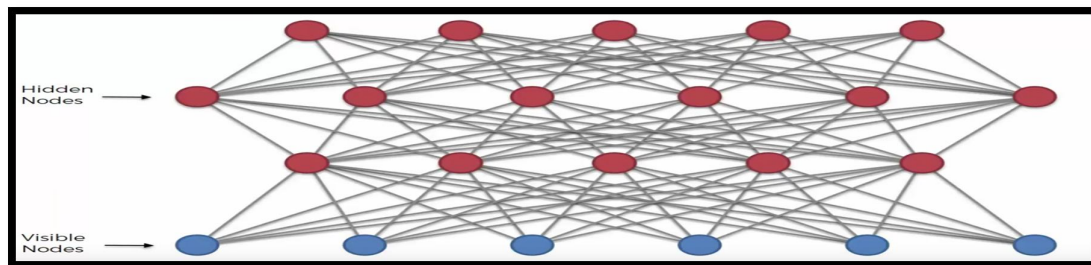


Fig 3. Architecture of DBN

C. Convolution Neural Network

CNN is used for image classification. To analyse the image, the image performs a convolution operation. First, convert the image into the form of a matrix representation, and then select the filter and perform the operation. To manage the quality of the image, they used the pooling function and maximum pooling. In CNN, data augmentation plays an important role in creating multiple copies of the same object by applying flipping, horizontal shifting, vertical shifting, and zooming. CNN is also used for text analysis and classification using the character or word embedding technique. They used a back-propagation algorithm to train the CNN model. Update the value of the filter based on the loss function. Wei et al. [15] Apply CNN to the proposed model. This study considers URL features to train the model and apply the character-level embedding technique. [16] The proposed model used CNN and the character embedding technique to detect a phishing attack. How to classify images shown in fig.4.

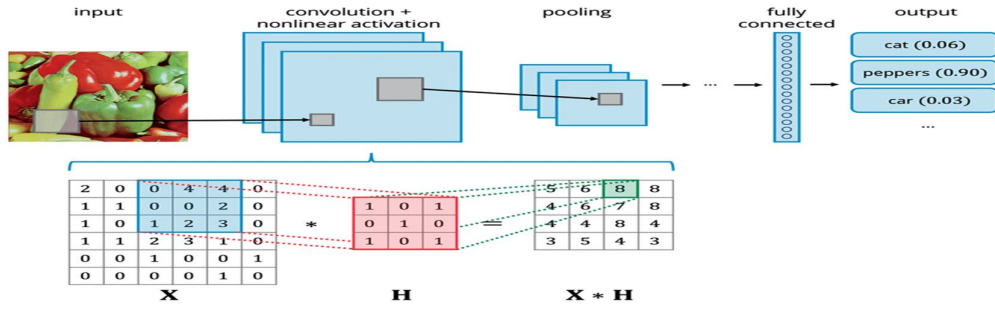


Fig 4. CNN Deep Learning Architecture

D. Recurrent Neural Network

RNN is used for semantic analysis, which checks whether it is a positive view or an opposing view. Analysis of the text in a sequential manner by RNN considers the current input and past output to produce the final output. The RNN takes the number of the word as an input matrix and then passes the first word to the hidden layer at the T1 time stamp. At the T2 time stamp, pass the next word and the output of the T1 time stamp. The detailed architecture of RNN is shown in fig5.

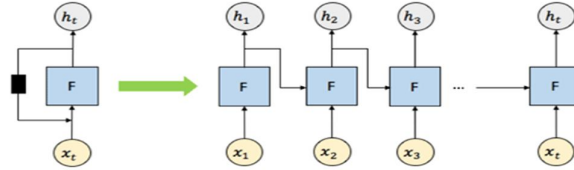


Fig 5. RNN Deep Learning Architecture

- F contains the function
- $x_1, x_2, x_3, \dots, x_n$ contains the input of words,
- The output is $h_1, h_2, h_3, \dots, h_n$. Calculated using this equation: $h_t = f(h_{t-1}, x_t), h_t = \tanh(w_{hh} h_{t-1} + w_{xh} x_t)$

To train the model using a back propagation algorithm, to assign the same weight to each input in the forwarding direction, but in the case of backward propagation, the value of the weight is changed. RNN has a vanishing gradient problem. [17] Alejandro Correa Bahnsen proposed a model using RNN to detect phishing websites. The accuracy is 98.7%. Consider the features of URLs and determine whether the URL is phishing or not.

E. Long Short-Term Memory

Barah Mohammed, Alanzi [18]. proposed a model using LSTM to detect phishing websites. LSTM is a special kind of RNN. It is used to resolve the problem of long-term dependency that occurred in RNN.

LSTM uses a memory cell, a forget gate, an input gate, and an output gate. The complete architecture of the LSTM model is shown in Fig. 6. Memory cells are based on remembering and forgetting content. In the forget gate, the previous (h_{t-1}) and new information (x_t) multiply with weight and pass to the sigmoid function. After that, add the information to the memory cell. The input gate is used to add data to the memory cell. The input gate uses the Tanh function that converts input in the form of +1 to -1. At the output, the gate takes the information from the memory cell and passes it to the next layer of LSTM.

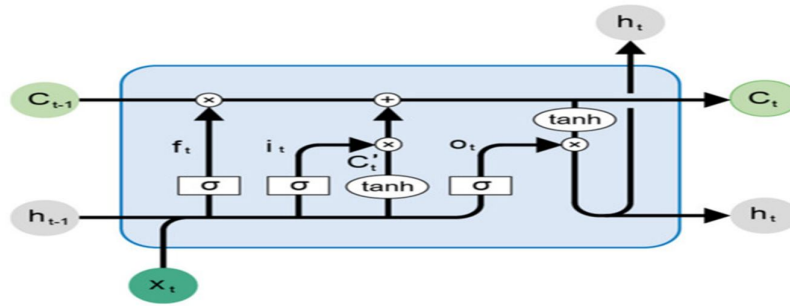
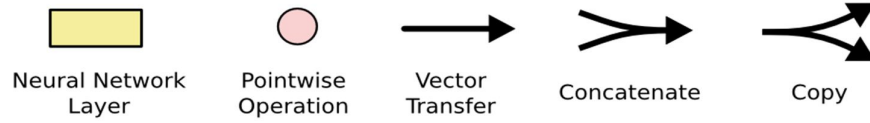


Fig.6 Architecture of LSTM Deep Learning model



An LSTM model is used to learn large-scale dependencies from the input sequence. It uses error gradients to calculate and combines iterative real-time learning with back propagation. [19].

F. Bidirectional-LSTM

BI-LSTM is an advanced version of LSTM. "BI-LSTM is the process of making any neural network have the sequence of information in both directions, backward (future to past) or forward (past to future)." [20]. In BI-LSTM, input can flow in either direction, forward or backward. Input in LSTM can only move in one direction, but with the help of BI-LSTM, the feature and past can be preserved. Deng, [20], Liang, Y. They proposed the model using bidirectional LSTM with 95.7 accuracies. The architecture of BI-LSTM is described in Fig. 7.

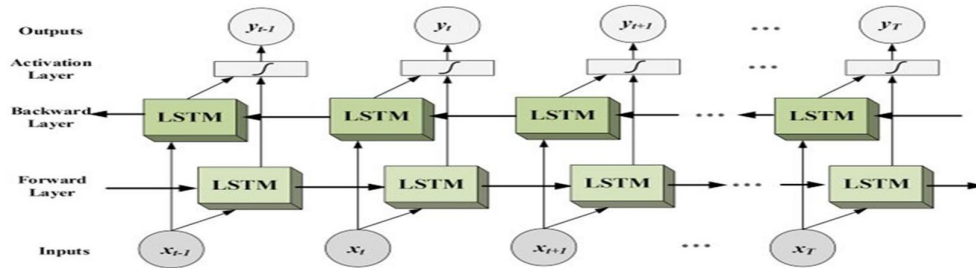


Fig 7. The Architecture of the BI-LSTM Deep Learning Model.

In Fig 7. Can see the flow of information from forward and backward directions. This model is used for NLP tasks. Guangxi Xu [22]. In which the self-attention Bi-LSTM neural network model is combined with ALBERT [22], a lightweight word vector model of BERT [23]. Compare the accuracy with a ML algorithm, but BI-LSTM gives a batter result as compared to others. The proposed model is 1% to 4% ahead of other models [23]. The proposed model adds ALBERT to get more accuracy.

G. Auto-encode

It is an unsupervised model. It is used to encode the data, pass it to the hidden layer, and decode the data. It is used to remove the dimension of the dataset. Reduce the size of data-driven applications. It can be used for image processing as well. Auto encoders can be used for several tasks, such as dimensionality reduction, effective feature learning, [23] nonlinear regression, etc. It is used to convert the input into the output with the least possible amount of derivation. It is used as an information compressor. [24] Sefer Kurnaz. They proposed a model using the deep auto encoder technique to classify phishing websites. Auto encoders are used to extract sensitive features and reduce the dimensions of features. For classification, use the softmax activation function. Collect the dataset from the UCI repository. [25] This paper proposed a model using an auto encoder and a denoising auto encoder (DAE). Therefore, to extract the complex features of the URL, the features space could be minimized significantly, thus reducing the phishing detection time. The architecture of the auto encoder is shown in Fig.8.

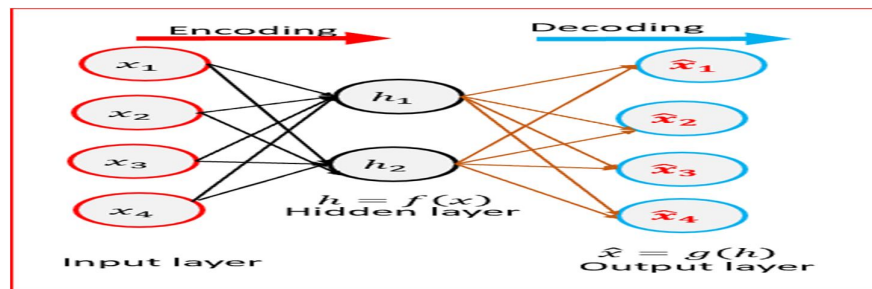


Fig.8 Architecture of the Auto encoder in Deep Learning

V. LEXICAL URL FEATURES OF URL DETECTION

To select the feature in deep learning, use the **sklearn library**, which provides many techniques used to extract the feature from the dataset. Some of them are described in detail:

- **It is used to remove all features with low variance.** It is used in unsupervised learning. Low variance contains constant values.
- **with feature selection:** In which to remove those values which are highly correlated. In which we remove highly correlated data from the training dataset as well as the testing dataset.
- **Mutual information in classification:-** Mutual information is based on information gain. It is used to find the dependency between variables. It returns that 0 means low dependence and 1 means high dependence between them. So, after applying mutual information in the classification, find those values which have low dependence between an independent and dependent variable and remove them from the dataset. To use the **Mutual_info_classify** library function. From sklearn, the python library used the **SelectBest()** method to dissect the best features that have the highest dependency between them. The library uses in Python is **Import mutual_info_classif from sklearn.feature_selection.**
- **Chi-square Test for Feature Selection:** It is used in univariate selection to select the best feature using the **SELECTKBBEST library** function in Python. The library used in python is "from sklearn. Feature-selection import select best, from sklearn.feature-selection import chi2. First, convert the dataset into the format of independent features; then, convert it into the format of dependent features. Apply it to the proposed model after then applying the chi-square test and finding the best feature on the basis of average values.

Feature Importance: This technique gives you a score for each feature of your data. The higher the score, the more relevant it is. The library used in Python is from **The Ensemble Import ExtraTreesClassifier is used.** Apoorva Joshi [25]. In which proposed the model using lexical features of URL and Machine Learning technique. To extract the feature, first understand the data, data preparation, data modeling, evaluation, deployment, and business understanding.

Data Understanding: -Include ~5 million URLs. Which was collected from Open phish, Alexa, and Internal FireEye.

Data Preparation :- In this phase, first take URL features and convert them into a numerical vector format. Used 23 features of URLs, which are described in table 2. To prepare the dataset with the help of NLTK (NLP Python package), mmh3 (Murmur Hash Python package), and urllib (Python library for parsing URLs) in Python.

Data Modeling: Used Random Forest and Decision Tree to propose the model to detect malicious URLs.

Suresh Dara [27]. This paper gives a detailed description of various feature extraction techniques used in DL. In this paper, input image feeds to MLP-based feature, MLP-based image and CNN model. These three methods collect the input image directly and apply the image to the individual neurons for classification.

Innocent Mbona[28]. Benford's law is used to predict the frequency of distribution of a significant leading digit. This study shows that features sharply obey Benford's law on a human dataset, whereas a small number of features violate Benford's law on a malicious bot dataset. Principal component analysis is productive for resolving high-dimensional imbalanced data problems. It is used to reduce the dimension to a low-dimensional value. Ala Mughald [30]. In this paper, we propose a model based on machine learning techniques. Find the accuracy of the proposed model's analysis based on three datasets. The first dataset contains 20 features of URLs. The second dataset contains 50 features about URLs. The third dataset contains 2500 ham and 500 spam emails. The first experiment shows that boosted decision trees give the best accuracy. The second experiment showed that the booster decision tree again gave the best accuracy. In the third expression, neural networks give high accuracy. The restriction of this study was the discovery of the predefined dataset. In terms of features, try to develop an automated tool to extract current features from the latest raw emails to enhance the accuracy of the proposed model.

VI. PERFORMANCE EVALUATION

Fig. 8 shows the accuracy of the existing models. All models have good accuracy. So, in this feature, we attempt to propose the model that provides the highest accuracy compared to the existing model.

VII. CONCLUSION

In recent years, phishing URLs have become a common issue. Many methodologies have been used to resolve the trouble of phishing attacks. All techniques have some advantages as well as some disadvantages. Deep

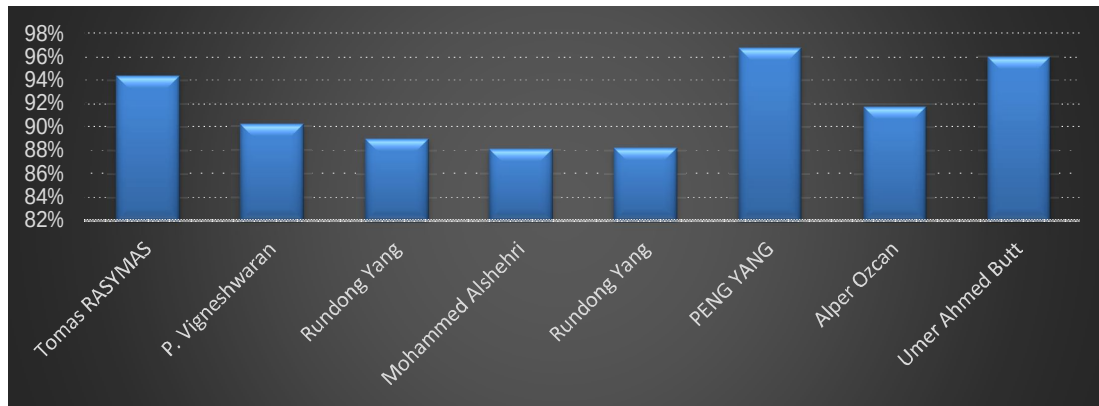


Fig 10. Analysis the accuracy of existing model

learning and machine learning are critical to developing a model to combat phishing URL attacks. This paper analyses various deep learning models using feature selection techniques. We analysed various techniques for feature selection (variance threshold, feature selection with correlation, mutual information in classification, and the Chi-square test for feature selection). In deep learning, we analyse the accuracy of an existing model. In this feature, we will demonstrate the model using the deep learning technique. CNN and BI-LSTM provide the highest accuracy in deep learning to resolve the problem of phishing URL attacks.

DECLARATIONS

“Conflict of interest the author have no relevant financial or non-financial interests to disclose.”

REFERENCES

- [1] Seok-Jun Bu, Hea-Jung Kim, Optimized URL Feature Selection Based on Genetic-Algorithm-Embedded Deep Learning for Phishing Website Detection. In Proceeding of the Electronics 2022,11,1090,https://doi.org/10.3390/electronics11071090.
- [2] Ali Aljofey, Qingshan Jiang, "An Effective Phishing Detection Model Based on Character-level Convolutional Neural Network from URL," proceeding by Electronics 2020, 9, 1514, DOI:-10.3390/electronics9091514.
- [3] Chidimma Opera, Bo Wei, Yingke, HTMLPhish: Enabling Phishing Web page Detection by Applying Deep Learning Techniques on HTML Analysis, international joint Conference on Neural Networks (IJCNN) 2020, DOI:-https://doi.org/10.1109/IJCNN48605.2020.9207707.
- [4] Tomas RASYMAS, Laurynas DOVYDAITIS, "Detection of Phishing URLs by Using Deep Learning Approach and Multiple Feature Combinations", In Proceeding by Modern Computing, Vol.8 (2020), No3,471-483, DOI:-https://doi.org/10.22364/bjmc.2020.8.3.06.
- [5] PENG YANG, GUANGZHEN ZHAO, "Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning," Published by IEEE, Volume 7, 2019, DOI:-10.1109/ACCESS.2019.2892066.
- [6] Hung Le and Quang Pham, "URLNet: Learning a URL Representation with Deep Learning for Malicious URL Detection", In Proceeding of the ACM Conference, Washing, DC, USAK, 2017, DOI:https://doi.org/10.475/123_4
- [7] P.vigneshwaran, A. Soumith Roy, "Multidimensional Features Driven Phishing Detection Based on Deep Learning," in a proceeding by the International Conference on Intelligent Emerging Methods of Artificial Intelligence & Cloud Computing, pp-362-369,2021. DOI:-https://doi.org/10.1007/978-3-030-92905-3_45.
- [8] Rundong Yang, Kangteng Zheng, "Predicting User Susceptibility to Phishing Based on Multidimensional Features", In proceeding Hindawi Computational Intelligence and Neuroscience volume 2022, Article ID 7058972,11 pages, DOI:-https://doi.org/10.1155/2022/7058972.
- [9] Mohammed Alshehri, Ahed Abugabah, "Character-level word encoding deep learning model for combating cyber threats in phishing URL detection," published by Computers and Electrical Engineering,2022.
- [10] Umer Ahmed Butt, Rashid Amin, "Cloud-based email phishing attack using machine and deep learning algorithm", in proceeding Complex & intelligent system,2022, DOI:-https://doi/10.1007/s40747-022-00760-3.
- [11] Alper Ozcan, Cagatay Catal, "A hybrid DNN-LSTM model for detecting phishing URLs", In a proceeding by Neural Comput & Applic (2021), DOI:https://doi.org/10.1007/s00521-021-06401-z.
- [12] Yi P, Guan Y, Zou F, Yao Y, Wang W, Zhu T (2018). Web phishing detection using a deep learning framework. Wireless Commun Mobile Comput 2018:4678746

- [13] Maneesh Kumar Verma, Shankar Yadav, Bhoopesh Kumar Goyal, Bakshi Rohit Prasad, and Sonali Agarawa, "Phishing Website Detection Using Neural Network and Deep Belief Network", Recent Findings in Intelligent Computing Techniques, Advances in Intelligent Systems and Computing 707, https://doi.org/10.1007/978-981-10-8639-7_30.
- [14] Abunadi, A., Akanbi, O., Zainal, A.: Feature extraction process: a phishing detection approach. In: Intelligent Systems Design and Applications (ISDA), 2013 13th International Conference on, pp. 331–335. IEEE (2013)
- [15] Wei W, Ke Q, Nowak J, Korytkowski M, Scherer R (2020) Accurate and fast URL phishing detector: a convolutional neural network approach. *Comput Netw* 178:107275
- [16] Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP (2020) An effective phishing detection model based on character level convolutional neural network from URL. *Electronics* 9:9.
- [17] Alejandro Correa Bahnsen; Eduardo Contreras Bohorquez; Sergio Villegas; "Classifying phishing URLs using recurrent neural networks", Proceeding of the 2017 APWG Symposium on Electronic Crime Research (eCrime), DOI: 10.1109/ECRIME.2017.7945048, IEEE
- [18] Barah Mohammed Alanzi1 and Diaa Mohammed Uliyan1, "Detection of Phishing Websites by Investigating Their URLs using LSTM Algorithm ", Published by IJCSNS International Journal of Computer Science and Network Security, VOL.22 No.5, May 2022.
- [19] Dargan S, Kumar M, Ayyagari MR, Kumar G (2020) A survey of deep learning and its applications: a new paradigm to machine learning. *Arch Comput Methods Eng* 27(4):1071–1092 Complete Guide To Bidirectional LSTM (With Python Codes) (analyticsindiamag.com)
- [20] Yuchen Liang, Jiangdong Deng, and Baojiang Cui, "Bidirectional LSTM: An Innovative Approach for Phishing URL Identification," Proceeding in Innovative Mobile and Internet Services in Ubiquitous Computing. Advances in Intelligent Systems and Computing, vol. 994. Springer, Cham. 2020.DOI:-https://doi.org/10.1007/978-3-030-22263-5_32
- [21] Guangxia Xu., Daiqi Zhou, and Jun Liu, "Social Network Spam Detection Based on ALBERT and Combination of Bi-LSTM with Self-Attention", Hindawi Security and Communication Networks Volume 2021, Article ID 5567991, 11 pages <https://doi.org/10.1155/2021/5567991>
- [22] Chathurdara Sri Nadith PathirageJun LiLing LiHong HaoWanquan LiuPinghe Ni, "Structural damage identification based on autoencoder neural networks and deep learning", Engineering Structures Volume 172, 1 October 2018, Pages 13-28,DOI:-<https://doi.org/10.1016/j.engstruct.2018.05.109>.
- [23] Sefer Kurnaz1, Wisam Gwad2, "Deep Auto-Encoder Neural Network for Phishing Website Classification," IJCSMC, Vol. 7, Issue. 3, March 2018, pg.68 – 72.
- [24] Samira Douzi, Meryem Amar, and Bouabid El Ouahidi, "Advanced Phishing Filter Using Autoencoder and Denoising Autoencoder," BDIOT2017, December 20–22, 2017, London, United Kingdom Association for Computing Machinery , DOI: 10.1145/3175684.3175690, December 2017.
- [25] Apoorva Joshi, Levi Lloyd, Paul Westin, "Using Lexical Features for Malicious URL Detection - A Machine Learning Approach", Cryptography and Security (cs.CR),2019, 29. <https://doi.org/10.48550/arXiv.1910.06277>
- [26] Saleem Raja, Vinodini R., Kavitha A., "Lexical features based malicious URL detection using machine learning techniques", 12th National Conference on Recent Advancements in Biomedical Engineering, 2021, Elsevier Ltd,DOI:-<https://doi.org/10.1016/j.matpr.2021.04.041>
- [27] Suresh Dara, Priyanka Tumma, "Feature Extraction By Using Deep Learning: A Survey", Proceedings of the 2nd International Conference on Electronics, Communication, and Aerospace Technology (ICECA 2018), IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-5386-0965-1
- [28] Innocent Mbona and Jan H.P. Eloff, "Feature selection using Benford's law to support detection of malicious social media bots", Information Sciences 582 (2022) 369–381, DOI:-<https://doi.org/10.1016/j.ins.2021.09.038>.
- [29] Ala Mughaid1 • Shadi AlZu'bi2 • Adnan Hnaif2 • Salah Taamneh1 • Asma Alnajjar1 • Esraa Abu Elsoud1, "An intelligent cyber security phishing detection system using deep learning techniques" Cluster Computing <https://doi.org/10.1007/s10586-022-03604-4>,2022.
- [30] APWG GA, Manning R(2020)Apwg phishing reports
- [31] Lastdrager ee(2014) Achieved a consensual definition of phishing based on a systematic review of the literature. *Crime Sci* 3(1):9