The International Journal of Engineering and Science (IJES)
|| Volume || 9 || Issue || 06 || Series I || Pages || PP 57-62 || 2020 ||
ISSN (e): 2319-1813 ISSN (p): 20-24-1805

# A Predictive Model for Phishing Attack Detection Using Machine Learning

Ashutosh Tiwari, Prof. Vipul Dalal
[1] *Computer Engineering, Vidyalankar Institute of Technology, Maharashtra, India.*
[2] *Computer Engineering, Vidyalankar Institute of Technology. Maharashtra, India.*

-----***ABSTRACT***-----
*These days, numerous enemy of phishing frameworks are being created to recognize phishing substance in online correspondence frameworks. In spite of the accessibility of hordes hostile to phishing frameworks, phishing proceeds with unabated because of lacking recognition of a zero-day assault, pointless computational overhead and high bogus rates. In spite of the fact that Machine Learning approaches have accomplished promising exactness rate, the decision and the exhibition of the component vector limit their successful location. Phishing is a typical assault on guileless individuals by making them to unveil their one of a kind data utilizing fake sites. In this work, an upgraded AI based prescient model is proposed to improve the effectiveness of against phishing plans. The prescient model comprises of Feature Selection Module which is utilized for the development of a successful element vector. These highlights are removed from the URL, website page properties and site page conduct utilizing the gradual segment-based framework to introduce the resultant component vector to the prescient model. The proposed framework utilizes CNN, KNN AND SVM which have been prepared on a 30-dimensional list of capabilities. AI is an incredible asset used to endeavor against phishing assaults.*

***KEYWORDS;-****Phishing, Phishing Websites, Detection, CNN,SVM,KNN.*

Date of Submission: 28-05-2020 Date of Acceptance: 14-06-2020

## I. INTRODUCTION

Phishing is the most unsafe crimes in the web. Since most by far of the customers go online to get to the organizations gave by government and cash related establishments, there has been a basic augmentation in phishing attacks for whatever length of time that very few years. Phishers started to pick up money and they are doing this as a productive business. Various procedures are used by phishers to ambush the feeble customers, for instance, advising, VOIP, criticized association and phony locales. It is definitely not hard to make counterfeit destinations, which looks like a bona fide site with respect to structure and substance. To be sure, the substance of these locales would be unclear from their certifiable destinations. The reasonfor making these destinations is to get private data from customers like record numbers, login id, passwords of charge and Mastercard, etc. Also, attackers request that security requests answer to acting like a noteworthy level wellbeing exertion providing for customers. Exactly when customers respond to those requests, they get easily captured into phishing attacks. Various investigates have been continuing to thwart phishing ambushes by different systems far and wide. Phishing attacks can be hindered by recognizing the locales and making consideration regarding customers to perceive the phishing destinations. Computer based intelligence counts have been one of the weighty techniques in recognizing phishing locales. In this assessment, various procedures for recognizing phishing locales have been discussed.

### 1.1 Machine Learning Algorithm

Two machine learning classification model K Nearest Neighbor and Support vector machine has been selected to detect phishing websites.

**K Nearest Neighbor Algorithm**

KNN is a viable administered learning strategy for some, issues including security techniques-nearest neighbor depends on the bunching of the components that have similar attributes; it chooses the class classification of a test model dependent on its k neighbor that is close to it. The estimation of k in the KNN relies upon the size of dataset and the sort of the grouping issue.

**Support Vector Machine Algorithm**

Support vector machine is another powerful algorithm in machine learning technology. In support vector machine algorithm, each data item is plotted as a point in n-dimensional space and support vector machine algorithm constructs separating line for classification of two classes, this separating line is well known as hyperplane.

Support vector machine seeks for the closest points called as support vectors and once it finds the closest point it draws a line connecting to them. Support vector machine then construct separating line which bisects and perpendicular to the connecting line. In order to classify data perfectlythe margin should be maximum. Here the margin is a distance between hyperplane and support vectors. In real scenario it is not possible to separate complex and nonlinear data, to solve this problem support vector machine uses kernel trick which transforms lower dimensional space to higher dimensional space.

### 1.2 Deep Learning-Based Detection

To address the previously mentioned absconds, some profound learning-based phishing site discovery arrangements have as of late become exposed because of the accomplishment in Natural Language Processing (NLP) accomplished by profound learning. A few inquiries about [19], [20], [21] concentrated on recognizing phishing URLs by utilizing the potential attributes of URLs. Conversely, different investigations [22] completely misused substance-based highlights or occasion-based highlights for phishing sites location. The essential distinction of our methodology with respect to the recently refered to profound learning based ones is that we consolidate the upsides of CNN and consideration based various leveled RNN to separate novel character level spatial and word-level worldly element portrayals of URLs naturally, which end up being helpful for the improvement of phishing identification execution.

## II. LITERATURE REVIEW

Creators in this paper[1] disclosed a novel way to deal with distinguish phishing sites utilizing AI calculations. They additionally looked at the precision of five AI calculations Decision Tree (DT), Random Forest (RF)[1], Gradient Boosting (GBM), Generalized Linear Model (GLM) and Generalized Additive Model (GAM)[1]. Exactness, Precision and Recall assessment techniques were determined for every calculation and thought about. Site qualities (30) are extricated with the assistance of Python and execution assessment finished with open source programming language R. Top three calculations to be specific Decision Tree, Random Forest and GBM execution were looked at in table. From the tables of exactness, review and execution, it is indicated that Random Forest calculation has given most noteworthy 98.4% precision, 98.59% review and 97.70% accuracy. In this paper creators [2] proposes a characterization mode[2]l so as to group the phishing assaults. This model includes highlight extraction from locales and characterization of site. In include extraction, 30 highlights have been taken from UCI Irvine AI vault informational index and phishing highlight extraction rules has been obviously characterized.So as to characterization of these highlights, Support Vector Machine (SVM), Naïve Bayes (NB) and Extreme Learning Machine (ELM)[2] were utilized. In Extreme Learning Machine (ELM), six actuation capacities were utilized and accomplished 95.34% precision than SVM and NB. The outcomes were acquired with the assistance of MATLAB. Creators [3] presents a way to deal with distinguish phishing email assaults utilizing characteristic language handling and AI. This is utilized to play out the semantic investigation of the content to identify malignant aim. A characteristic Language Processing (NLP) strategy is used to parse each sentence and secures the semantic positions of words in the sentence in association with the predicate. Considering the activity of each word in the sentence, this procedure perceives whether the sentence is a request or a request. Directed machine learning[3] is utilized to produce the boycott of vindictive sets. Creators characterized calculation SEAHound[3] for recognizing phishing messages and Netcraft Anti-Phishing Toolbar is utilized to confirm the legitimacy of a URL. This calculation is actualized with Python contents and dataset Nazario phishing email set is utilized. Aftereffects of Netcraft and SEAHound[3] are thought about and acquired exactness 98% and 95% separately.This outcome exhibits that semantic information is a strong pointer of social structuring.

Another methodology by creators [4] proposes include determination calculations to diminish the parts of dataset to get higher request execution [4]. It likewise contrasted and other information mining grouping calculations and results acquired. Dataset for phishing sites was taken from UCI AI repository[4]. From the results, it is seen that some arrangement procedures increase the execution; some of them decay the execution with diminished segment. Bayesian Network, Stochastic Gradient Descent (SGD), lazy.K.Star, Randomizable Filtered Classifier, Logistic model tree (LMT) and ID3 (Iterative Dichotomiser)[4] are helpful for decrease phishing dataset and Multilayer Perception, JRip, PART, J48[4], Random Forest and Random Tree calculations are not important for the reduced phishing dataset. Lazy.K.Star got 97.58% exactness with 27 diminished highlights. This investigation is gotten with the assistance of WEKA programming.

Creators [5]proposed a model with answer for perceive phishing destinations by using URL distinguishing proof procedure using Random Forest calculation. Show has three phases, specifically Parsing, Heuristic Classification of information, Performance Analysis [5]. Parsing isutilized to dissect include set.

Dataset accumulated from Phishtank. Out of 31 highlights just 8 highlights are considered for parsing. Arbitrary woods strategy got exactness level of 95%.

Creators [6] proposed an adaptable sifting choice module to separate highlights consequently with no particular master information on the URL space utilizing neural system model. In this methodology creators utilized all the characters remembered for the URL strings and tally byte esteems. They not just check byte esteems and furthermore cover portions of neighboring characters by moving 4-bits. They implant mix data of two characters showing up consecutively and tallies how frequently each worth shows up in the first URL string and accomplishes a 512 measurement vector. Neural system model tried with three analyzers Adam, AdaDelta and SGD. Adam was the best enhancer with exactness 94.18% than others. Creators additionally infer that this model exactness is higher than the recently proposed complex neural system topology.

In this paper creators [7] made a similar report to identify pernicious URL with traditional AI procedure – strategic relapse utilizing bigram, profound learning methods like convolution neural system (CNN) and CNN long momentary memory (CNN-LSTM)[7] as engineering. The dataset gathered from Phishtank, OpenPhish for phishing URLs and dataset MalwareDomainlist, MalwareDomains were gathered for malevolent URLs. Because of correlation, CNN-LSTM acquired 98% precision. In this paper creators utilized TensorFlow[7] in conjuction with Keras[7] for profound learning engineering.

Creators in this paper [8] likewise proposed decreased component choice model to recognize phishing sites. They utilized Logistic Regression and Support Vector Machine (SVM)[8] as arrangement techniques to approve the element determination strategy. 19 highlights decreased from 30 site highlights have been chosen and utilized for phishing recognition. The LR and SVM computations execution was reviewed subject to exactness, review, f-measure and precision. Study shows that SVM calculation accomplished best execution over LR calculation.

In this paper creators [9] proposed a phishing identification model to distinguish the phishing execution successfully by utilizing mining the semantic highlights of word implanting, semantic element and multi-scale factual features[9] in Chinese website pages. Eleven highlights wereextricated and classified into five classes to secure measurable highlights of site pages. AdaBoost, Bagging, Random Forest and SMO[9] are utilized to actualize learning and testing the model. Real URLs dataset acquired from DirectIndustry web guides and phishing information was gotten from Anti-Phishing Alliance of China. As per study, just semantic highlights very much distinguished the phishing destinations with high detection[9] effectiveness and combination model accomplished the best execution recognition. This model is extraordinary to Chinese site pages and it has reliance in certain language.

This paper [10] proposes a productive method to distinguish phishing URL sites by utilizing c4.5 choice tree approach. This method extricates highlights from the locales and figures heuristic qualities. These qualities were given to the c4.5 choice tree algorithm[10] to decide if the site is phishing or not. Dataset is gathered from PhishTank and Google. This procedure incorporates two stages in particular pre-handling stage and identification phase[10]. In which highlights are extricated dependent on rules in pre-handling stage and the highlights and their regarded qualities were inputted to the c4.5 calculation and acquired 89.40% precision.

Creators [11] in this paper made an expansion to Google Chrome to distinguish phishing sites content with the assistance of AI calculations. Dataset UCI-Machine Learning Repository utilized and 22 highlights were separated for this dataset. Calculations kNN, SVM and Random Forest were picked for exactness, recall,f1-score and precision examination. Irregular Forest got a best score and HTML, JavaScript, CSS[11] utilized for executing chrome augmentation alongside python. This augmentation is having a downside of proclaimed pernicious site list which is expanding each day.

This paper [12] approaches a structure to extricate highlights adaptable and straightforward with new procedures. Information is gathered from PhishTank[12] and genuine URLs from Google[12]. To get the content properties C# programming and R writing computer programs were utilized. 133 highlights were gotten from the dataset and outsider specialist co-ops. CFS subset based and Consistency subset-based element selection[12] strategies utilized for highlight determination and examined with WEKA instrument. Guileless Bayes and Sequential Minimal Optimization (SMO)[12] calculations were thought about for execution assessment andSMO is favored by the creator for phishing discovery than NB.

Another heuristic highlights identification strategy by creators [13] clarifies about the element of URL, for example, PrimaryDomain, SubDomain, PathDomain and positioning of site, for example, PageRank, AlexaRank, AlexReputation to recognize the phishing sites. Dataset utilized from PhishTank and trial is splitted into 6 stages through MYSQL, PHP with 10 testing datasets. The proposed model contains two stages. In Phase I site highlights were removed and in Phase II six estimations of heuristic are determined. As indicated by creators, if heuristic worth is closest to one, the site is considered as genuine and in the event that it is closest to zero, at that point the site is questioned as phishing site. Root Mean Square Error (RMSE)[13] is utilized to ascertain precision and got 97% exactness.

In this paper creator [14] presents a phishing URL recognition framework relies upon URL lexical investigation named PhishScore. This methodology depends on intra-URL relatedness[14][18]. This relatedness mirrors the relationship into part of the URL Right around 12 site features expelled from a lone URL are used to incorporate AI calculations to recognize phishing URLs. This test results precision of 94.91%.

## III. PROPOSED SYSTEM

Our proposed to discover whether the URL is phishing or nonphishing utilizing CNN, Support Vector Machine and KNN. Support Vector Machine is additionally called a Support Vector arrange which is a managed learning model that related with learning calculation which is utilized to look at the information for order and relapse. We have a lot of preparing test and every one of this test has a place with any of three classes. Bolster Vector Machine calculation distributes another guide to one or different classifications.

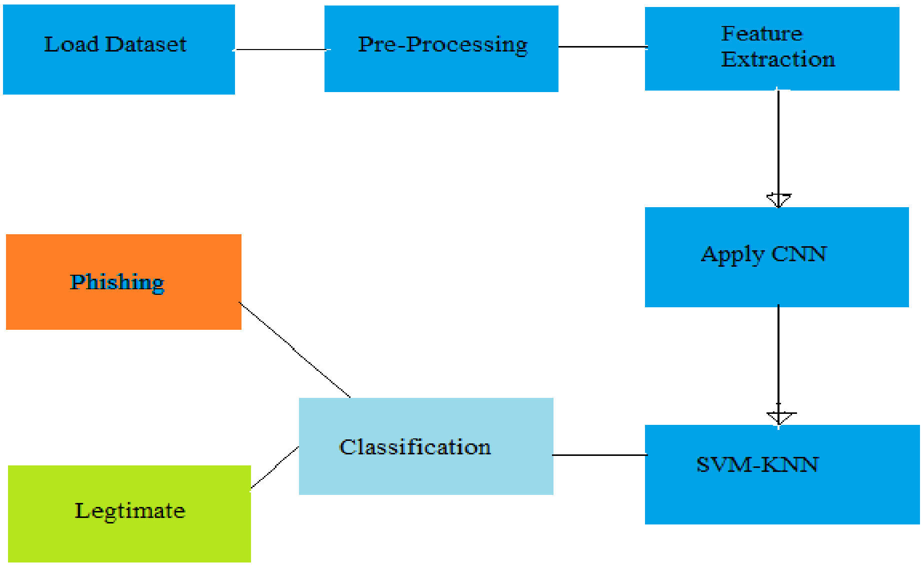

**Fig.1 Proposed Architecture**

## IV. RESULT VIEW

| Algorithm | Feature | Accuracy |
|---|---|---|
| Random Forest | 30 | 79.86 |
| NB | 30 | 72.69 |
| Proposed(CNN+SVM+KNN) | 30 | 92.68 |

**Table 1: Outline of Algorithms used to detect Phishing Website URLs**

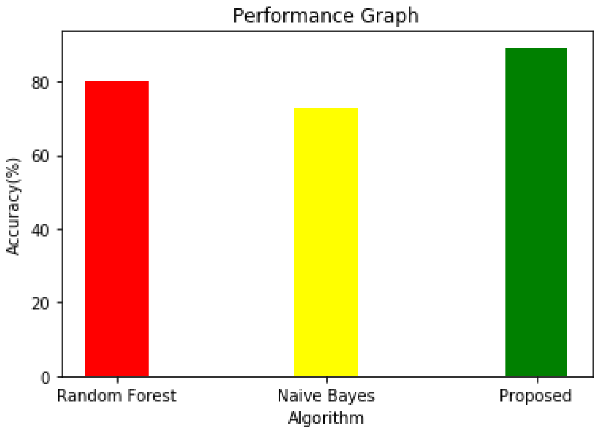

**Fig.2: Performance analysis on the basis of table1.**

In this experimental approach, we obtained the performance of the proposed approach in term of accuracy concerning the subset of the feature set. This was done by building the predictive model to consider only the subset of the selected features. Specifically, the subset of the feature set was selected from the 30 feature categories used in this work. These features were selected to measure their impact on the evaluation process and determine if there is any contributory influence or complementing effects on the other features that are not considered.Our experiment on phishing dataset indicates that CNN+SVM+KNN returned 92.68% ,Random Forest returned 79.86 while NB gave 72.69%

## V. CONCLUSION

The primary thought in this paper is to explore how existing phishing feature datasets can be adequately coordinated into a successful countermeasure. To accomplish this, the positioning of various component classes from surviving writing was utilized to choose our proposed feature vector

In light of this reason, the proposed approach separated a few highlights from the URL, site page properties and website page conduct utilizing recurrence appraisal examination. At long last, a gradual development model was utilized to sort out the highlights and its related AI models into a framework for adaptability and sensibility. The methodology was assessed utilizing probes the classifiers comprising of KNN and SVM with a dataset comprising of 2541 phishing pages and 25,000 authentic pages. The outcomes demonstrate a pleasant runtime of less 2,000 ms and assessment measurements comprising of 99.96 True Positives, 99.96 True Negatives, 0.04 False Positive and 0.04 False Negatives. From these outcomes, the proposed approach presents a predominant enemy of phishing plan when contrasted and other existing methodologies under the given test conditions.

The graph databases and relational database both performed well. In general, graph databases performed better when objective tests were performed. The Implementation shows that graph databases retrieve the results of the set of predefines query faster than relational databases.Alsograph databases are more flexible than relational databases as we can add new relationships to graph databases without the need to restructure the schema again.

## REFERENCE

**[1].** J. Shad and S. Sharma, “A Novel Machine Learning Approach to Detect Phishing Websites Jaypee Institute of Information Technology,” pp. 425–430, 2018.

[2]. Y. Sönmez, T. Tuncer, H. Gökal, and E. Avci, “Phishing web sites features classification based on extreme learning machine,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018

[3]. T. Peng, I. Harris, and Y. Sawa, “Detecting Phishing Attacks Using Natural Language Processing and Machine Learning,” Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018–Janua, pp. 300–301, 2018.

[4]. M. Karabatak and T. Mustafa, “Performance comparison of classifiers on reduced phishing website dataset,” 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, pp. 1–5, 2018.

[5]. S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, “A New Method for Detection of Phishing Websites: URL Detection,” in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[6]. K. Shima et al., “Classification of URL bitstreams using bag of bytes,” in 2018 21st Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN), 2018, vol. 91, pp. 1–5.

[7]. A. Vazhayil, R. Vinayakumar, and K. Soman, “Comparative Study of the Detection of Malicious URLs Using Shallow and Deep Networks,” in 2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018, 2018, pp. 1– 6.

[8]. W. Fadheel, M. Abusharkh, and I. Abdel-Qader, “On Feature Selection for the Prediction of Phishing Websites,” 2017 IEEE 15th Intl Conf Dependable, Auton. Secur. Comput. 15th Intl Conf Pervasive Intell. Comput. 3rd Intl Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr., pp. 871–876, 2017

[9]. X. Zhang, Y. Zeng, X. Jin, Z. Yan, and G. Geng, “Boosting the Phishing Detection Performance by Semantic Analysis,” 2017.

[10]. L. MacHado and J. Gadge, “Phishing Sites Detection Based on C4.5 Decision Tree Algorithm,” in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.

[11]. A. Desai, J. Jatakia, R. Naik, and N. Raul, “Malicious web content detection using machine leaning,” RTEICT 2017 - 2nd IEEE Int. Conf. Recent Trends Electron. Inf. Commun. Technol. Proc., vol. 2018–Janua, pp. 1432–1436, 2018.

[12]. M. Aydin and N. Baykal, “Feature extraction and classification phishing websites based on URL,” 2015 IEEE Conf. Commun. NetworkSecurity, CNS 2015, pp. 769–770, 2015.

[13]. L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen, “A novel approach for phishing detection using URL-based heuristic,” 2014 Int. Conf. Comput. Manag. Telecommun. ComManTel 2014, pp. 298–303, 2014.

[14]. S. Marchal, J. Francois, R. State, and T. Engel, “PhishScore: Hacking phishers’ minds,” Proc. 10th Int. Conf. Netw. Serv. Manag. CNSM 2014, pp. 46–54, 2015.

[15]. A. A. Ahmed and N. A. Abdullah, “Real time detection of phishing websites,” 7th IEEE Annu. Inf. Technol. Electron. Mob. Commun. Conf. IEEE IEMCON 2016, 2016

[16]. X. Zhang, Y. Zeng, X. B. Jin, Z. W. Yan, and G. G. Geng, “Boosting the phishing detection performance by semantic analysis,” in Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017, 2018, vol. 2018– Janua, pp. 1063–1070.

[17]. Dr.D.Akila, Dr.C. Jayakumar, "Acquiring Evolving Semantic Relationships for WordNet to Enhance Information Retrieval", International Journal of Engineering and Technology, Volume 6, November 5, pp. 2115-2128, 2014.

[18]. . D.Akila,S.Sathya, G.Suseendran, “Survey on Query Expansion Techniques in Word Net Application”, Journal of Advanced Research in Dynamical and Control Systems, Vol.10(4), pp.119-124, 2018.

[19]. A. C. Bahnsen, E. C. Bohorquez, S. Villegas, J. Vargas, and F. A. González, “Classifying phishing urls using recurrent neural networks,” in Electronic Crime Research (eCrime), 2017 APWG Symposium on. IEEE, 2017, pp. 1–8.

[20]. H. Le, Q. Pham, D. Sahoo, and S. C. Hoi, “Urlnet: Learning a url representation with deep learning for malicious url detection,” arXiv preprint arXiv:1802.03162, 2018.

[21]. J. Saxe and K. Berlin, “expose: A character-level convolutional neural network with embeddings for detecting malicious urls, file paths and registry keys,” arXiv preprint arXiv:1702.08568, 2017.

[22]. B. Athiwaratkun and J. W. Stokes, “Malware classification with lstm and gru language models and a character-level cnn,” in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 2482–2486.

Ashutosh Tiwari, et. al. "A Predictive Model for Phishing Attack Detection Using Machine Learning." *The International Journal of Engineering and Science (IJES)*, 9(6), (2020): pp. 57-62.