iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VII **Month of publication:** July 2023

DOI: https://doi.org/10.22214/ijraset.2023.54854

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

Phishing Website Prediction Using Gradient Boosting Classifier

Dr. M Prasad[1], Ansifa Kouser M[2]
[1]*Associate Professor, Department of Master of Computer Application, BIET, Davangere*
[2]*Fourth Semester Student, Department of MCA, BIET, Davangere*

Abstract: ***Phishing is a widespread tactic used to trick gullible people into disclosing their personal information by using bogus websites. Phishing website URLs are designed to steal personal data, including user names, passwords, and online financial activities. Phishers employ websites that resemble those genuine websites both aesthetically and linguistically. Utilizing anti-phishing methods to identify phishing is necessary to stop the rapid advancement of phishing techniques as a result of advancing technology. A strong tool for thwarting phishing assaults is machine learning. Attackers frequently use phishing because it is simpler to fool a victim into clicking a malicious link that looks authentic than to try to get past a computer's security measures. The malicious links within the message body are intended to appear to go to the spoofed company utilising that company's logos and other genuine information. In the method that is being presented, machine learning is used to create a revolutionary approach for detecting phishing websites. Gradient Boosting Classifier is the model we utilised in our suggested strategy to identify phishing websites based on aspects of URL significance. By extracting and comparing different characteristics between legitimate and phishing URLs, the suggested method uses gradient boosting classifier to identify phishing URLs. The studies' findings demonstrate that the suggested approach successfully identifies legitimate websites from bogus ones in real time.***
Keywords: URL, Phishing and Gradient Boosting Classifier.

I. INTRODUCTION

Artificial intelligence is a new innovative science that reviews and creates hypotheses, strategies, procedures, and applications that recreate, grow and broaden human knowl-edge.ML is an arm of artificial intelligence and it is analogous to (and frequently overlap with) computational measurements, that also concentrates on making predictions with the use of PCs. Machine leaning has solid relationship with scientific improvement, which tells methods, hypothesis and utilization regions to the field. ML is sometimes, in a while combined with data mining, but the data mining subfield focuses more on preparatory information investigation and is called as unsupervised learning.ML can like wise be unsupervised and be utilized to learn and set up pattern profiles for various entities and then used to find important anomalies.

Cyber security is a set of innovations and procedures intended to secure PCs, networks, projects and information from assaults and unapproved access, modification, or annihilation .A system security framework comprises of a system assurance frame work and furthermore a PC protection framework. Every one of these frame works incorporates firewalls, antivirus programming, and intrusion detection system (IDS). IDSs help find, decide and distinguish unapproved system conduct , for instance, use, replicating, change and annihilation.

There are three important kind of network analysis for Intrusion detection system misuse based, also known as anomaly-based, signature-based, and hybrid.

1) Misuse based detection strategies mean to distinguish realized attacks by utilizing the marks of these attacks.
2) Anomaly-based methods study the typical system and conduct and distinguish anomalies as deviations from ordinary behavior.
3) Hybrid detection conflates anomaly and misuse detection .To expand the rate of detection of accepted intrusions and to decrease the rate of false positives of unknown attacks.

The applications of machine learning (ML) methods in cyber security is rising than ever before. Beginning from IP traffic categorization, separating malicious traffic for intrusion detection, Machine learning is the one of the best answers that can impact against zero-day attacks. New exploration is being done by utilization of measurable traffic characteristics and ML techniques . The word phishing was introduced in the year 1987 . Phishing is an online thievery that robs an individual’s private data and identity data. A sort of extortion where the assailant gets complete access to other individual’s private data.

Because of increase in the phishing attacks, numerous results are proposed which generates a solution to the issue. To build a framework which guarantees a solution against the phishing attack, there are several ways. Various other methods for detecting phishing attack are there like black list, Fuzzy rule-based, white list-based, cantina-based, machine learning based, Heuristic and image-based approaches . There are several other studies that talks about a variety of methods and techniques to detect the different types of phishing attacks .Phishing sites looks to be like a genuine website and several individuals have problem in recognizing such websites. Few anti-phishing techniques are in built in some of the browsers.

II. LITERATURE SURVEY

H. Huang et al., (2009) proposed the frameworks that distinguish the phishing utilizing page section similitude that breaks down universal resource locator tokens to create forecast preciseness phishing pages normally keep its CSS vogue like their objective pages.

S. Marchal et al., (2017) proposed this technique to differentiate Phishing website depends on the examination of authentic site server log knowledge. An application Off-the- Hook application or identification of phishing website. Free, displays a couple of outstanding properties together with high preciseness, whole autonomy, and nice language-freedom, speed of selection, flexibility to dynamic phish and flexibility to advancement in phishing ways.

Mustafa Aydin et al. proposed a classification algorithm for phishing website detection by extracting websites' URL features and analyzing subset based feature selection methods. It implements feature extraction and selection methods for the detection of phishing websites.

The extracted features about the URL of the pages and composed feature matrix are categorized into five different analyses as Alpha- numeric Character Analysis, Keyword Analysis, Security Analysis, Domain Identity Analysis and Rank Based Analysis. Most of these features are the textual properties of the URL itself and others based on third parties services.

In the existing system they have used Logistic Regression, Multinomial Naive Bayes, and XG Boost are the machine learning methods that are compared. The Logistic Regression algorithm outperforms the other two. The model is preprocessed in the proposed system, the words are tokenized, and stemming is performed.

Data Processing is the process of converting or encoding data for easy machine transfer. The accuracy of Logistic Regression is 96.63 percent, and the overall comparison is presented.

Advancements in information technology often task users with complex and consequential privacy and security decisions. A growing body of research has investigated individuals' choices in the presence of privacy and information security tradeoffs, the decision-making hurdles affecting those choices, and ways to mitigate such hurdles. The article provides a multi-disciplinary assessment of the literature pertaining to privacy and security decision making. And focuses on research on assisting individuals' privacy and security choices with soft paternalistic interventions that nudge users toward more beneficial choices. The article discusses potential benefits of those interventions, highlights their shortcomings, and identifies key ethical, design, and research challenges.

III.SYSTEM DESIGN

System design thought as the application of theory of the systems for the development of the project. System design defines the architecture, data flow, use case, class, sequence and activity diagrams of the project development.

A. System Architecture

The architecture of the system is as shown in fig 1; the URLs to be classified as legitimate or phishing is fed as input to the appropriate classifier. Then classifier that is being trained to classify URLs as phishing or legitimate from the training dataset uses the pattern it recognized to classify the newly fed input.

The features such as IP address, URL length, domain, having favicon, etc. are extracted from the URL and a list of its values is generated. The list is fed to the classifiers such as KNN, kernel SVM, Decision tree and Random Forest classifier. These models' performance is then evaluated and an accuracy score is generated. The trained classifier using the generated list predicts if the URL is legitimate or phishing.

The list contains values 1, 0 and -1 if the features exist, not applicable and if the features doesn't exist respectively. There are 30 features being considered in this project

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VII Jul 2023- Available at www.ijraset.com

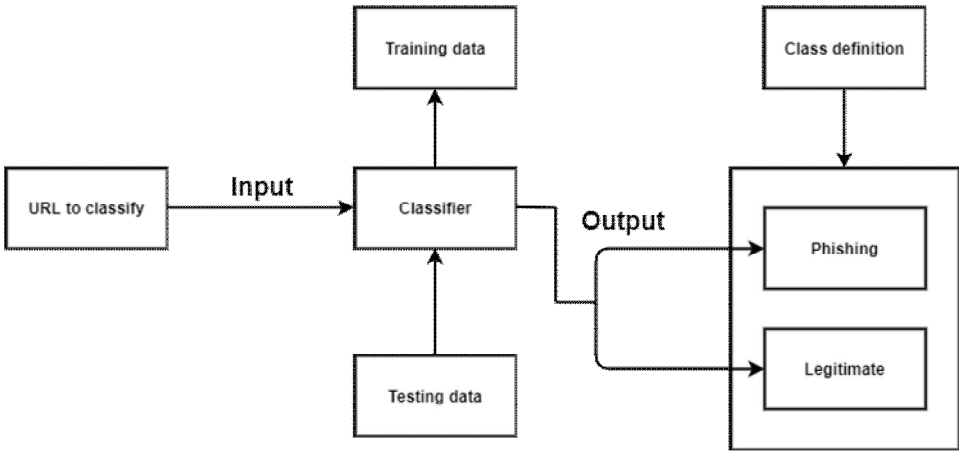


Fig. 1 Architecture Diagram

B. Dataflow Diagram

DFD level 1 gives a more detailed explanation of the Context diagram. The high-level process of the Context diagram is broken down into its subprocesses. The DFD level 1 of the system is depicted in fig 2. The Level 1 DFD takes a step deep by including the processes involved in the system such as feature extraction, splitting of dataset, building the classifier, etc. and hence gives a more detailed vision of the system.:

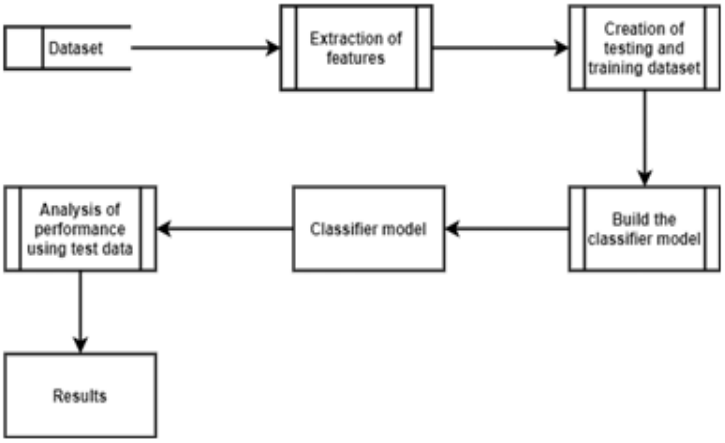


Fig. 2 Sequence Diagram

C. Implementation

1) *Data Collection:* In the first module developing the data collection process. This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that get. There are several techniques to collect the data, like web scraping, manual interventions. The dataset is referred from the popular dataset repository called kaggle. The following is the dataset link for the Detection of Phishing Websites Using Machine Learning.
2) *Data Preparation:* Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.). Randomize data, which erases the effects of the particular order in which the collected and/or otherwise prepared our data. Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis. Split into training and evaluation sets.
3) *Model Selection:* The Gradient Boosting Classifier machine learning algorithm. An accuracy of training Accuracy 98.9% so this algorithm is implemented.

4) *Gradient Boosting Classifier Algorithm:* While studying machine learning have come across this term called Boosting. And the most misinterpreted term in the field of Data Science. The principle behind boosting algorithms is first built a model on the training dataset, then a second model is built to rectify the errors present in the first model. When the target column is continuous, the use of Gradient Boosting Regressor whereas when is a classification problem, and the use of Gradient Boosting Classifier. The only difference between the two is the "Loss function". The objective here is to minimize this loss function by adding weak learners using gradient descent. And is based on loss function hence for regression problems, have different loss functions like Mean squared error (MSE) and for classification, and have different for e.g log-likelihood.

D. Understand Gradient Boosting

1) Step -1 The first step in gradient boosting is to build a base model to predict the observations in the training dataset. For simplicity take an average of the target column and assume that to be the predicted value and the loss function will be minimum so this value will become the prediction for the base model.
2) Step-2 The next step is to calculate the pseudo residuals which are (observed value – predicted value) The predicted value is the prediction made by the previous model.
3) Step-3: we will build a model on these pseudo residuals and make predictions.Need to minimize these residuals and minimizing the residuals will eventually improve the model accuracy and prediction power.
4) Step- 4 In this step find the output values for each leaf of decision tree. That means there might be a case where 1 leaf gets more than 1 residual, hence need to find the final output of all the leaves. TO find the output can simply take the average of all th numbers in a leaf, doesn't matter if there is only 1 number or more than 1.

Accuracy on test set: An accuracy of 97.6% on test set.
Saving the Trained Model: To take the trained and tested model into the production-ready environment, the first step is to save it into a .h5 or .pkl file using a library like pickle.

- Pickle installed in environment.
- Import the module and dump the model into.pkl file

IV.RESULTS

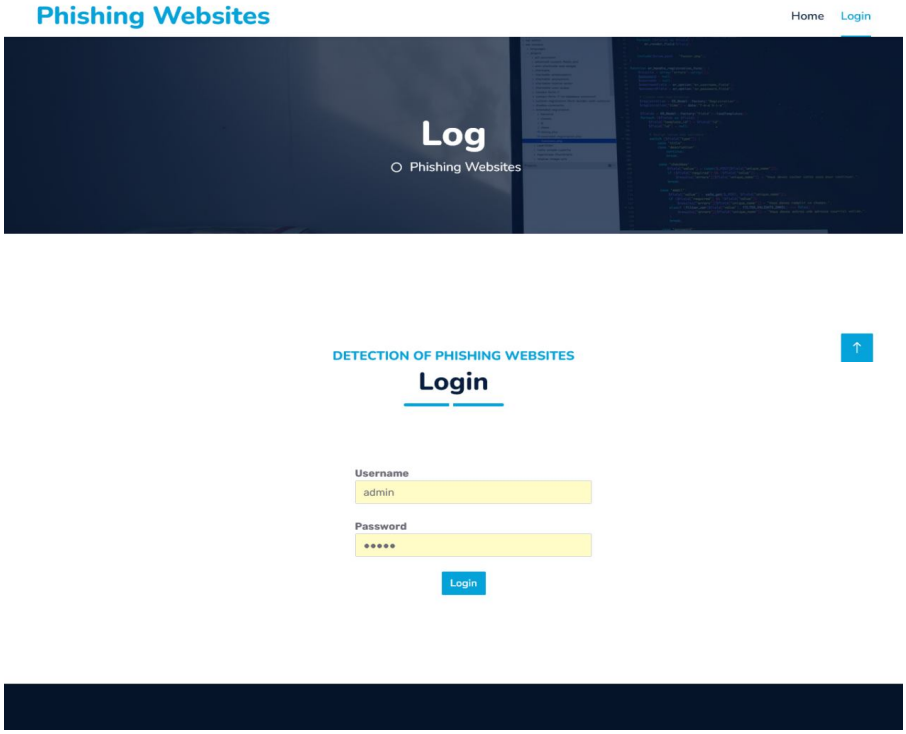


Fig. 3 Login Page

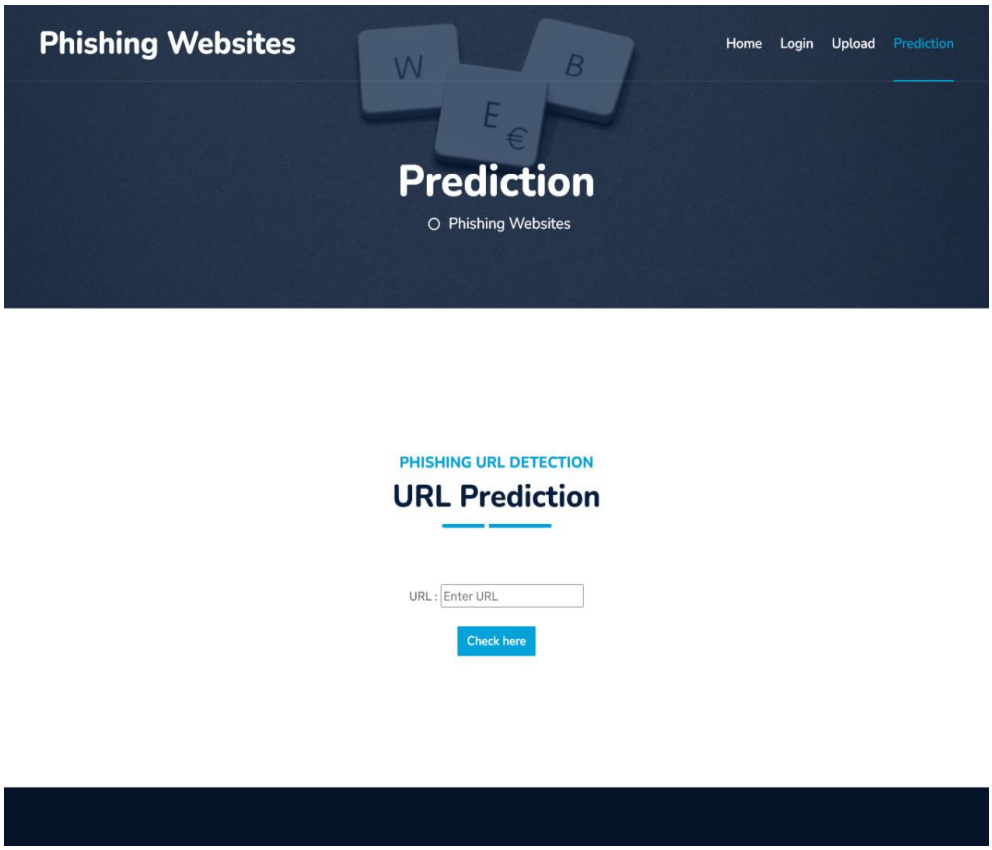


Fig. 4 URL uploading page

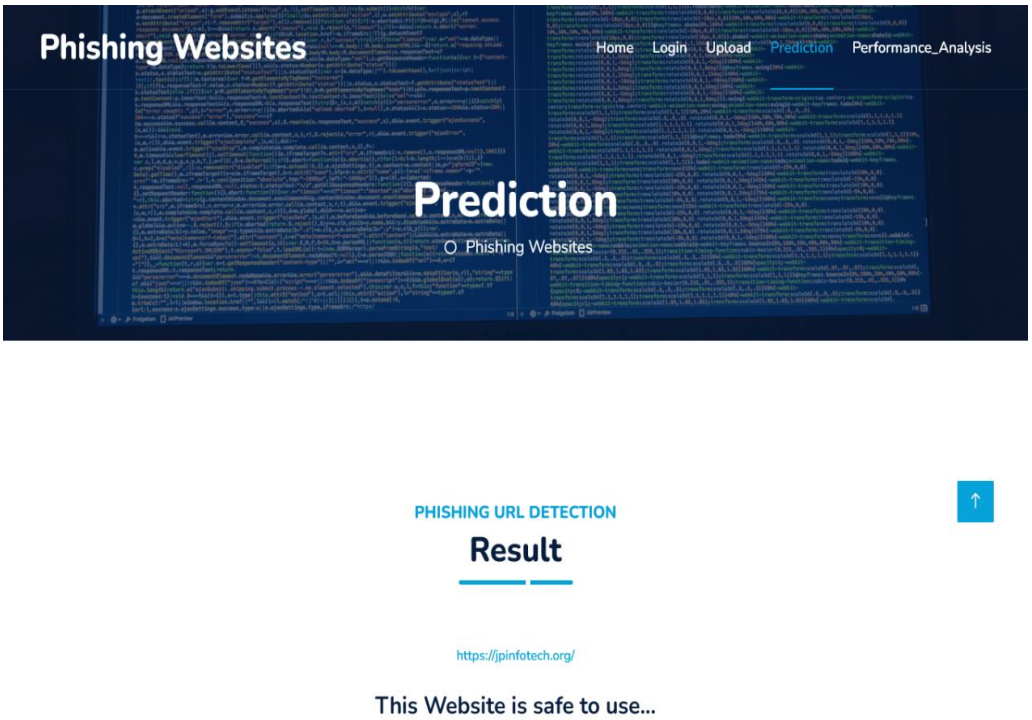


Fig. 5 Uploaded URL is predicted as safe

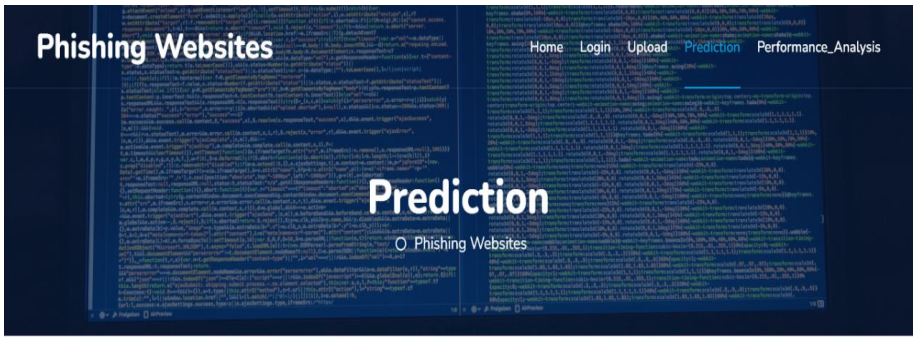


Fig. 6 Uploaded URL is predicted as unsafe

V.CONCLUSION

The work is remarkable that a good anti-phishing system should be able to predict phishing attacks in a reasonable amount of time. Accepting that having a good anti-phishing gadget available at a reasonable time is also necessary for expanding the scope of phishing site detection. The current system merely detects phishing websites using Gradient Boosting Classifier. 97% detection accuracy using Gradient Boosting Classifier with lowest false positive rate is achieved.

REFERENCES

[1] Chengshan Zhang, Steve Sheng, Brad Wardman, Gary Warner, Lorrie Faith Cranor, Jason Hong. Phishing Blacklists: An Empirical Study In: CEAS 2009: Proceedings of the 6th Conference on Email and Anti-Spam, Mountain View, California, USA, July 16-17, 2009.

[2] Andrew Jones, Mahmoud Khonji, Youssef Iraqi, Senior Member A Literature Review on Phishing Detection 2091-2121 in IEEE Communications Surveys and Tutorials, vol. 15, no. 4, 2013. 2013.

[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Many Understanding and Assisting Users' Online Choices with Nudges for Privacy and Security 50(3), Article No. 44, ACM Computing Surveys, 2017.

[4] Helena Matute, Mara M. Moreno-Fernández, Fernando Blanco, Pablo Garaizar I'm looking for phishers. To combat electronic fraud, Internet users' sensitivity to visual deception indicators should be improved. pp.421-436 in Computers in Human Behavior, Vol.69, 2017.

[5] F.J. Overink, M. Junger, L. Montoya. Preventing social engineering assaults with priming and warnings does not work. pp.75-87 in Computers in Human Behavior, Vol.66, 2017. 2017.

[6] M. El-Alfy, El-Sayed M. Probabilistic Neural Networks and K-Medoids Clustering are used to detect phishing websites. The Computer Journal, 60(12), pp.1745-1759, published in 2017.

[7] Shuang Hao, Luca Invernizzi, Yong Fang, Christopher Kruegel, Giovanni Vigna. Cheng Huang, Shuang Hao, Luca Invernizzi, Yong Fang, Christopher Kruegel, Giovanni Vigna. Gossip: Detecting Malicious Domains from Mailing List Discussions Automatically pp. 494-505 in Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security (ASIA CCS 2017), Abu Dhabi, United Arab Emirates, April 2-6, 2017.

[8] Gonzalo Nápoles, Rafael Falcon, Koen Vanhoof, Mario Köppen. Frank Vanhoenshoven, Gonzalo Nápoles, Rafael Falcon, Koen Vanhoof, Mario Köppen. Machine learning algorithms are used to detect dangerous URLs. The 2016 IEEE Symposium Series on Computational Intelligence (SSCI 2016) was held on December 6-9, 2016.

[9] Hillary Sanders, Joshua Saxe, Richard Harang, Cody Wild A Deep Learning Approach to Detecting Malicious Web Content in a Fast, Format-Independent Way. pp. 8-14 in Proceedings of the 2018 IEEE Symposium on Security and Privacy Workshops (SPW 2018), San Francisco, CA, USA, August 2.

[10] Jie Wu, Longfei Wu, Xiaojiang Du Phishing Attacks on Mobile Computing Platforms: Effective Defense Schemes 6678-6691 in IEEE Transactions on Vehicular Technology, vol. 65, no. 8, 2016.

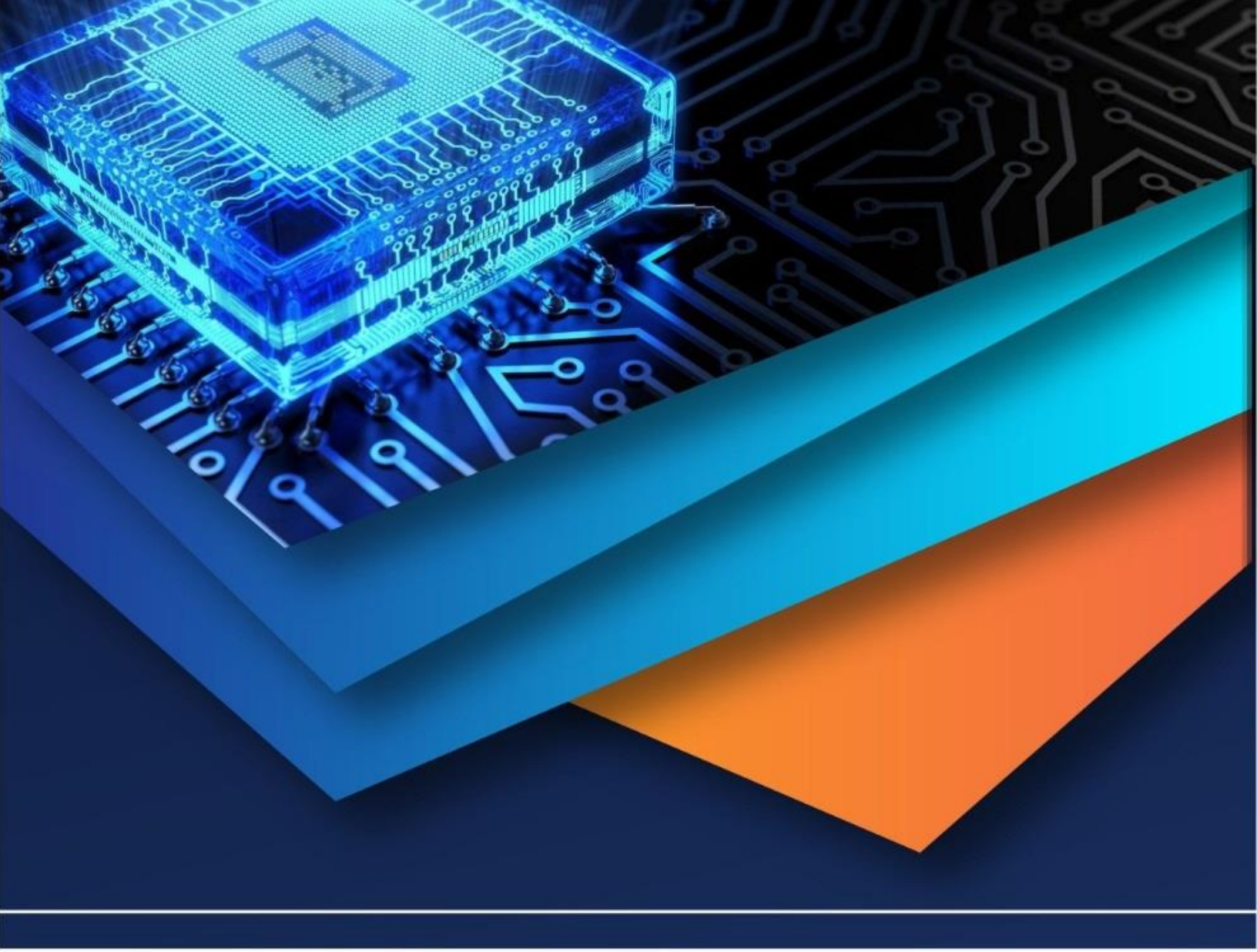
[11] Ilango Krishnamurthi, R. Gowtham A system for detecting phishing websites that is both thorough and effective. pp. 23-37 in Computers & Security, Vol. 40, 2014.

[12] Lorrie Cranor, Guang Xiang, Jason I. Hong, Carolyn Penstein Rosé CANTINA+: A Phishing Web Site Detection Framework with a Feature-Rich Machine Learning Framework. Article No. 21 in ACM Transactions on Information and System Security, 14(2), 2011.

[13] Chengcheng Ye, Erzhou Zhu, Dong Liu, Feng Liu, Futian Wang, Xuejun Li An Effective Phishing Detection Model Using Neural Networks and Optimal Feature Selection In: Proceedings of the IEEE International Symposium on Parallel and Distributed Processing with Applications, 16th IEEE International Symposium on Parallel and Distributed Processing with Applications 781-787, Melbourne, Australia, December 11-13, 2018. (ISPA 2018).

[14] Systematization of Knowledge (SoK): ASystematic Review of Software- Based Web Phishing Detection, by Zuochao Dou, Issa Khalil, Abdallah Khreishah, Ala Al-Fuqaha, and Mohsen Guizani, IEEE Communications Surveys & Tutorials, 2017.

[15] Detection and analysis of drive-by-download assaults and malicious javascriptcode," Proceedings of the 19th International Conference on World Wide Web, pp. 281-290, 2010. Marco Cova, Christopher Kruegel, Giovanni Vigna



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY