iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53910

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 VI Jun 2023- Available at www.ijraset.com

Machine Learning Approach for Detection of Phishing Website

Nikita Pawar[1], Dr. Pritish Tijare[2]
[1]*ME Student, Computer Science and Engineering, Sipna COET, Amravati, Maharashtra, India*
[2]*Professor, Computer Science and Engineering, Sipna COET, Amravati, Maharashtra, India*

Abstract: ***Phishing attacks are the easiest way to get sensitive information from innocent users. The purpose of phishing is to obtain critical information like the private information of users. Phishing sites pretend to be relevant sites and it is more difficult to distinguish these sites. It is one of the greatest threats that every individual and management has ever faced. URLs are referred to as web pages through which users locate pieces of information on the Internet. This paper focuses on phishing website detection which is important for the prevention of our sensitive information. There are different types of websites with different features. Therefore, we need to use a specific set of features on websites to protect against phishing. A machine learning model has been proposed to detect phishing websites. To detect phishing sites, we proposed the machine learning model. The essential objective of this paper is to classify the various algorithm of machine learning by extracting and analyzing various features of legitimate and phishing URLs.***
Keywords: Phishing Website, URL Detection, Machine Learning.

I. INTRODUCTION

As the Internet grows and expands, many of our occupations are now conducted online, along with e-commerce, business, social media, and banking, which increases the probability of online crime. Therefore, securing the World Wide Web is becoming more and more important [1].

Phishing is the fraud of deceiving a trusted person into an electronic connection to obtain all kind of information like person usernames, passwords, and information about credit card numbers. This is usually done through spoofing emails or instant messaging and often encourages consumers to enter personal information on a fake website that looks exactly like the real thing. The result is an information security breach that leads to a breach of private data, where the victim could potentially lose money or other assets. Internet users are exposed to various cyber threats including personal information theft, identity theft, and financial loss. Current research has generated a variety of techniques for detecting phishing websites including supervised machine learning. The supervised include all kinds of supervised machine learning methods (such as Support Vector Machines (SVM), Naive Bayes, Random Forest, etc.) [7]. The Machine learning approach prefers efficient classification of dataset. To increase the efficiency of the classification process, a specific processing technique is implemented. After Dataset training the machine is analyzed with several test datasets [2].

II. LITERATURE REVIEW

The author [1] used different machine learning classification algorithms to detect phishing websites and compare the resulting training dataset, since phishing and legitimate websites are separated in the dataset, and concluded that in the test dataset, 6,118 phishing websites and 3,610 legitimate websites were successfully detected from 9,879 data sets and a combined prediction rate of 98% was determined. There have been several studies focused on detecting phishing websites

The author [3] proposed a framework using a machine learning approach by using supervised machine learning as well as unsupervised machine learning and used the feature selection method, to analyze and reduce the redundancy of data which irrelevant or unnecessary in a data set.

The author [4] used various classifier models in machine learning and validate the model by using the feature selection method. According to this author, the highest accuracy rate achieved by the SVM algorithm by using feature based extraction with SVM.

The author [5] used RF and DT algorithms along with a feature extraction process for URLs and corresponding binary values for indicating whether the website is phishing or not. The author did this with a feature extraction method based on the IP address in the URL, @ symbol in the URL, prefix or suffix, HTTP tokens in the URL, and URL redirection.

©IJRASET: All Rights are Reserved

Table I: Related Work

Author Name	Classification Algorithm	Accuracy
Malak Aljabri, et.al. [1]	i] Support Vector Machine ii] Random Forest	99.8 99.7
Mohammad Nazmul Alam, et.al. [3]	i] Random Forest ii] Decision Tree	96.0 91.0
Uday Bhaskar Penta et al [4]	i] Support Vector Machine ii] Naive Bayes iii] K-Nearest Neighbors	94.5 91.6 89.8

III. PROBLEM ANALYSIS

Phishing websites are fake websites that can be built as legitimate websites to trick other people into stealing their important personal information like bank account information, social security number, and password. It will lead to information security breaches by stealing sensitive data resulting in financial loss to the victim. In short, it is an internet scam or a top criminal. Therefore, evaluating or detecting phishing sites requires an intelligent model to detect and detect suspicious characteristics associated with phishing sites. The main question addressed in this study is how to enhance user authentication on the website. Specifically, the goal here is to develop an aggregate model that will be used. To predict if a website is phishing or legitimate and to measure the accuracy of phishing site detection [6].

IV. PROPOSED WORK

In this paper, we analyze and classification of various machine learning algorithms.

A. *Process of Phishing Website*

1) Data collection: Dataset contains the 32 features. It is important for feature selection for the analysis of the dataset based on this feature we can make assumptions about whether the website trustable or not
2) Training Dataset: The training dataset required a component of a machine learning model and also it will help for accurate prediction. An evaluation of the model.
3) Testing Dataset: The test data set measures the accuracy and efficiency of the algorithm it gives the objective and evaluation of the model.

B. *Process of Machine Learning Technique:*

1) *Raw Dataset*: Select the complete raw dataset which contains the 32 features.
2) *Pre-processing module*: It processes the preparing the raw data. Separated the values (CSV) file and generate in the form of (1,-1,0).
3) *Processed Dataset*: After pre-processing the dataset will divide into testing and training datasets.
4) *Train/Test Dataset*: The dataset will be split into 80%-20.
5) *Classification of Algorithm*: The classification of the algorithm helps to categorize the data for detection of the website. After the training and testing process, the model gives accuracy.
6) *Result*: It detects the URL and shows whether it is safe or unsafe.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Based on the data the Accuracy, Precision, Recall, and f1score measures are the evaluation measures used to evaluate the classifier's performance.

- *Accuracy*: Accuracy is the ratio of the absolute value of the expected class to total readings.
  Accuracy = (TP + TN) / (TP + FP + TN + FN) [3]
- *Precision*: Precision is the ratio of the expected positive measure to the expected total positive measure.
  Precision =TP / (TP + FP) [3]
- *Recall*: Recall is the ratio of a correctly predicted positive measure to the total true score measure.
  Recall =TP / (TP + FN) [3]
- *F1 Score*: The F1 score is nothing but precision and recall weighted average. Therefore, this set contains both false positives and false negatives to ensure a balance between accuracy and memory.
  F1 Score = 2*((precision*recall)/(precision+recall)) [3]

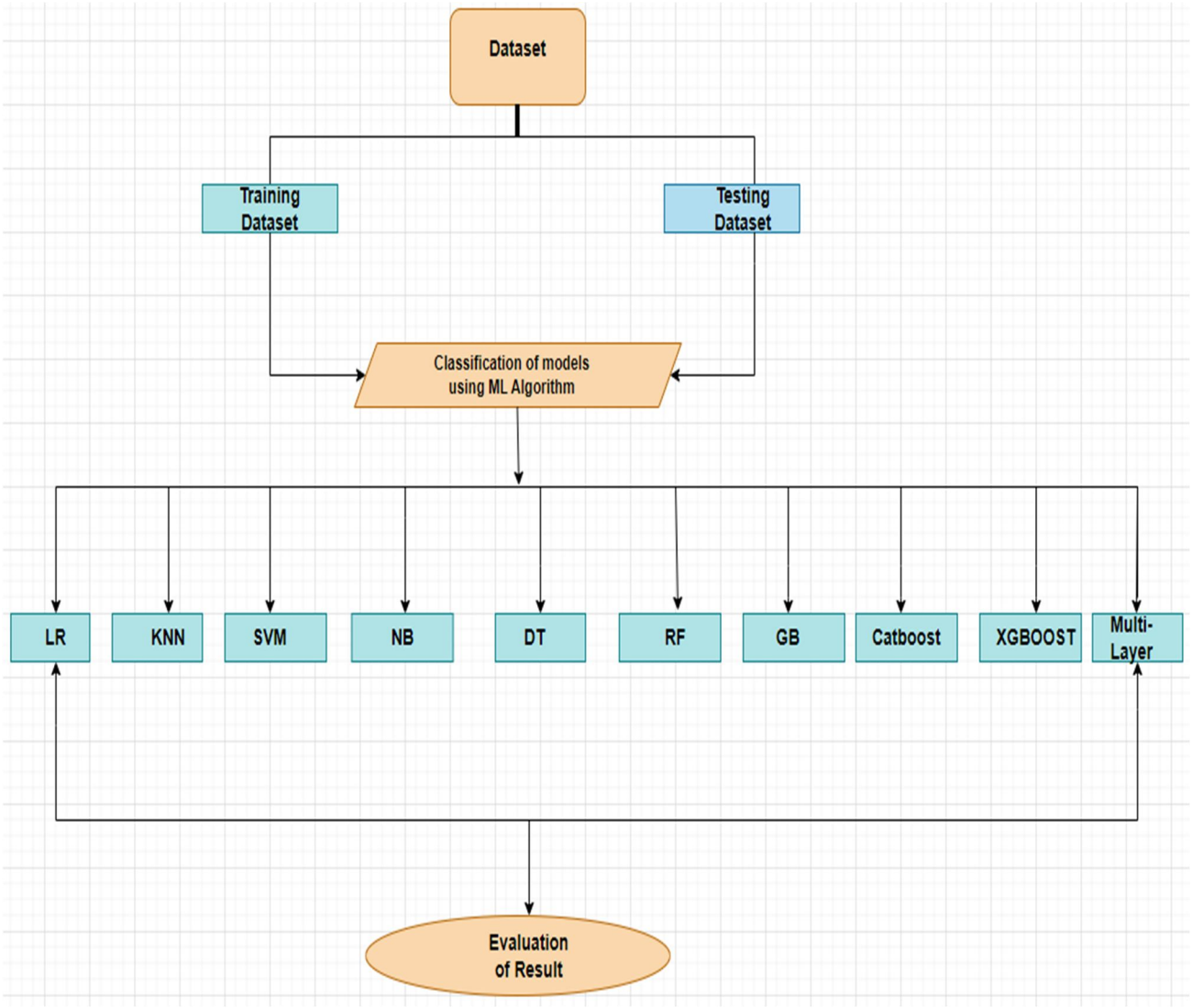


Figure:1 Flowchart of Proposed Work

The workflow of the project by using phishing detection is given in Figure 1, which provides the overview of the initial steps followed by applying a machine learning algorithm on the dataset. We classify our models by using various algorithms of machine learning like Logistic Regression, K-nearest neighbors, Support vector machine, Naive Bayes, Decision Tree, Random Forest, Gradient Boost, XG-Boost, Cat-Boost, and Multi-layer.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

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

V. RESULTS

A. *Classification of Algorithm*

1) *Logistic Regression:* We work on LR which shows the relationship between dependent and independent variables. The ratio between the variables is always balanced. We use LR on our dataset and according to that we test and train the dataset and find accuracy.

	precision	recall	f1-score	support
-1	0.94	0.91	0.92	976
1	0.93	0.95	0.94	1235
accuracy			0.93	2211
macro avg	0.93	0.93	0.93	2211
weighted avg	0.93	0.93	0.93	2211

Screenshot 1: Training and testing classification report of LR

2) *K- Nearest Neighbors:* The KNN assumes that the new dataset and existing data set is near about same and extract the feature of both dataset and according to that it gives the output.

	precision	recall	f1-score	support
-1	0.95	0.95	0.95	976
1	0.96	0.96	0.96	1235
accuracy			0.96	2211
macro avg	0.96	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

Screenshot 2: Training and testing classification report of KNN

3) *Support Vector Machine:* The relation between the given values and existing values will be straight. The prediction of our dataset and our existing data set must be the same.

	precision	recall	f1-score	support
-1	0.97	0.94	0.96	976
1	0.96	0.98	0.97	1235
accuracy			0.96	2211
macro avg	0.97	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

Screenshot 3: Training and testing classification report of SVM

4) *Naive Bayes:* In NB some features are independent on other features as well as some features are dependent of other features.

	precision	recall	f1-score	support
-1	0.97	0.94	0.96	976
1	0.96	0.98	0.97	1235
accuracy			0.96	2211
macro avg	0.97	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

Screenshot 4: Training and testing classification report of NB

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

5) *Decision Tree:* In DT we divide the phishing URL and non-phishing URL. The DT of each and every node denotes the features.

	precision	recall	f1-score	support
-1	0.96	0.96	0.96	976
1	0.97	0.96	0.97	1235
accuracy			0.96	2211
macro avg	0.96	0.96	0.96	2211
weighted avg	0.96	0.96	0.96	2211

Screenshot 5: Training and testing classification report of DT

6) *Random Forest:* Whatever different subset will be created in our training dataset we will apply RF inside it. According to its features, whether it is phishing or not will be decided.

	precision	recall	f1-score	support
-1	0.96	0.96	0.96	976
1	0.97	0.97	0.97	1235
accuracy			0.97	2211
macro avg	0.97	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

Screenshot 6: Training and testing classification report of RF

7) *Gradient Boost:* We created the dataset and used Gradient Boost when we will take our dataset to the next level the error will be less than before. It works sequentially. When we take value to the next step, the timing is between will check the speed of reading the data will be checked, hence it is called the learning rate.

	precision	recall	f1-score	support
-1	0.99	0.96	0.97	976
1	0.97	0.99	0.98	1235
accuracy			0.97	2211
macro avg	0.98	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

Screenshot 7: Training and testing classification report of Gradient Boost

8) *Cat-Boost:* When we have two kinds of datasets similar and non-similar and when we arrange the model that time we use Cat-Boost so that whatever error we have is removed in the first step itself.

	precision	recall	f1-score	support
-1	0.98	0.96	0.97	976
1	0.97	0.98	0.98	1235
accuracy			0.97	2211
macro avg	0.97	0.97	0.97	2211
weighted avg	0.97	0.97	0.97	2211

Screenshot 8: Training and testing classification report of Cat-Boost

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

9) *XG-Boost:* Whatever the model made after the boosting we check the accuracy of training and testing of that model.

```
XGBoost Classifier : Accuracy on training Data: 0.987
XGBoost Classifier : Accuracy on test Data: 0.969

XGBoost Classifier : f1_score on training Data: 0.988
XGBoost Classifier : f1_score on test Data: 0.973

XGBoost Classifier : Recall on training Data: 0.993
XGBoost Classifier : Recall on test Data: 0.993

XGBoost Classifier : precision on training Data: 0.984
XGBoost Classifier : precision on test Data: 0.984
```

Screenshot 9: Training and testing classification report of XG-Boost

10) *Multi-Layer:* It divides the dataset into linear form, it breaks the restrictions and classifies the dataset.

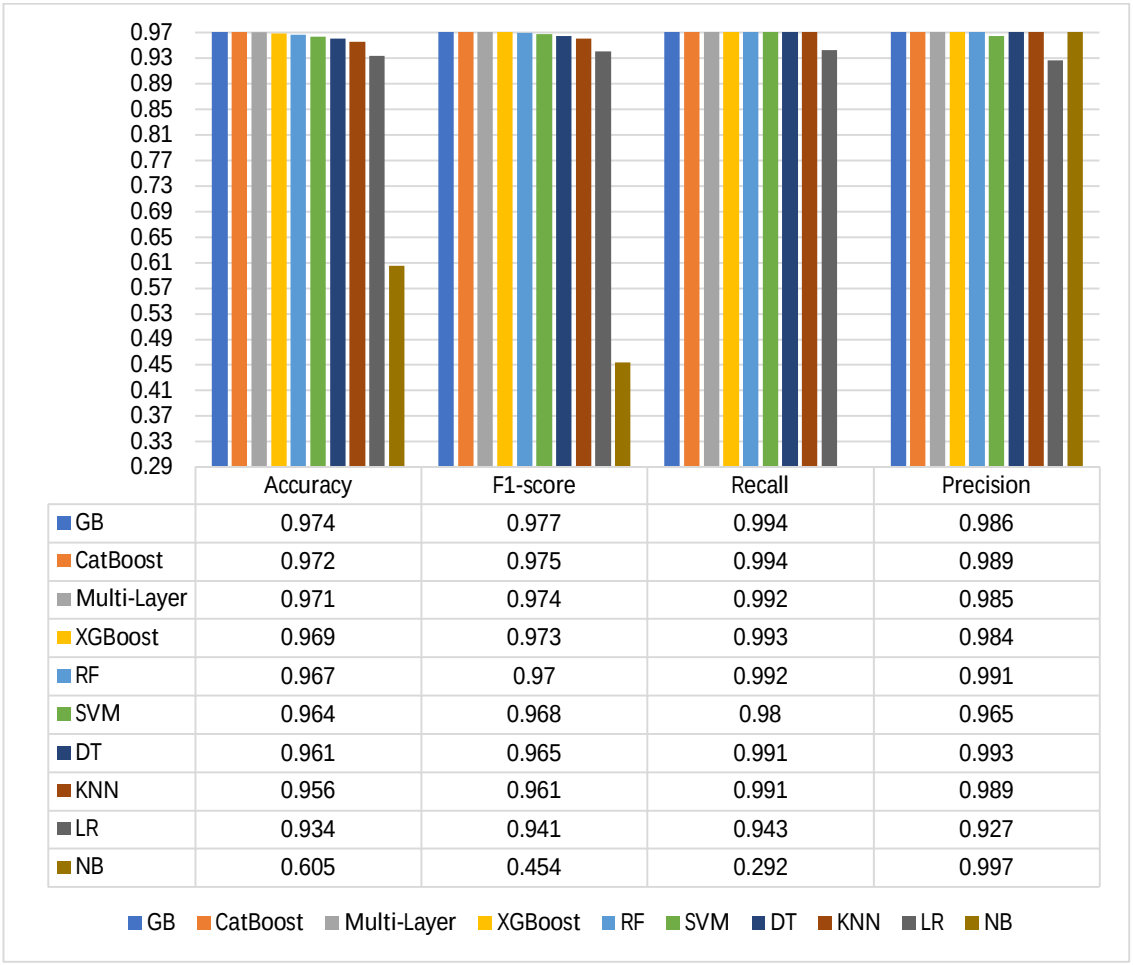
```
Multi-layer Perceptron : Accuracy on training Data: 0.987
Multi-layer Perceptron : Accuracy on test Data: 0.971

Multi-layer Perceptron : f1_score on training Data: 0.989
Multi-layer Perceptron : f1_score on test Data: 0.989

Multi-layer Perceptron : Recall on training Data: 0.992
Multi-layer Perceptron : Recall on test Data: 0.982

Multi-layer Perceptron : precision on training Data: 0.985
Multi-layer Perceptron : precision on test Data: 0.967
```

Screenshot 10: Training and testing classification report of Multi-layer



	Accuracy	F1-score	Recall	Precision
GB	0.974	0.977	0.994	0.986
CatBoost	0.972	0.975	0.994	0.989
Multi-Layer	0.971	0.974	0.992	0.985
XGBoost	0.969	0.973	0.993	0.984
RF	0.967	0.97	0.992	0.991
SVM	0.964	0.968	0.98	0.965
DT	0.961	0.965	0.991	0.993
KNN	0.956	0.961	0.991	0.989
LR	0.934	0.941	0.943	0.927
NB	0.605	0.454	0.292	0.997

Graph 1: Comparative Classification Report

The above graph 1 is the comparative classification report is a representation of all classifier models using a machine learning algorithm. The Gradient Boost achieved a high accuracy that is 0.974, f1_score is 0.977, recall is 0.994, and precision is 0.986 as compared to the other models, so we can say that this one is perfect. The NB classifier achieved a lower accuracy that is 0.605, f1_score is 0.454, recall is 0.292, and precision is 0.997. So we cannot consider that the NB is the perfect one.

VI. CONCLUSION

Based on machine learning the detection of phishing websites is proposed in this paper. A standard dataset used for the machine learning algorithm. We developed a model that can be particularly utilized in figuring out whether or not the website is both phishing or non-phishing. We applied algorithms like LR, KNN, SVM, NB, DT, RF, GB, Cat-Boost, XG-Boost, and Multi-Layer Perception to the classification model that has been developed. Therefore we achieved the highest accuracy of the Gradient boosting classifier model. Thus it can be concluded that the GB achieved highest accuracy and gave good result. This paper is intended to be useful to the reader and give a review analysis of these methods of the proposed system can find safe and unsafe websites.

REFERENCES

[1] Malak Aljabri, Hanan S. Altamimi, Shahd A. Albelali, Maimunah Al-Harbi, Haya T. Alhuraib, Najd K. Alotaibi, Amal A. Alahmadi, Fahd Alhaidari, Rami Mustafa A. Mohammad, Khaled Salah “Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions”, IEEE Access, Volume 10, doi:10.1109/ACCESS.2022.3222307,pp. 121395- 121417,2022.

[2] Anuja Bhosale, Gayatri Gadas , Muskan Chavan , Neha Pandhare , Seema Hadke,” Detection of Phishing Website Using Machine Learning”, International Journal of Advanced Research in Computer and Communication Engineering, Volume 11, Issue 6, doi: 10.17148/IJARCCE.2022.11695,pp. 490-494, June 2022.

[3] Mohammad Nazmul Alam, Dhiman Sarma, Farzana Firoz Lima, Ishita Saha, Rubaiath-E- Ulfath, “Phishing Attacks Detection using Machine Learning Approach”,IEEE Xplore, pp. 1173-1179,2020.

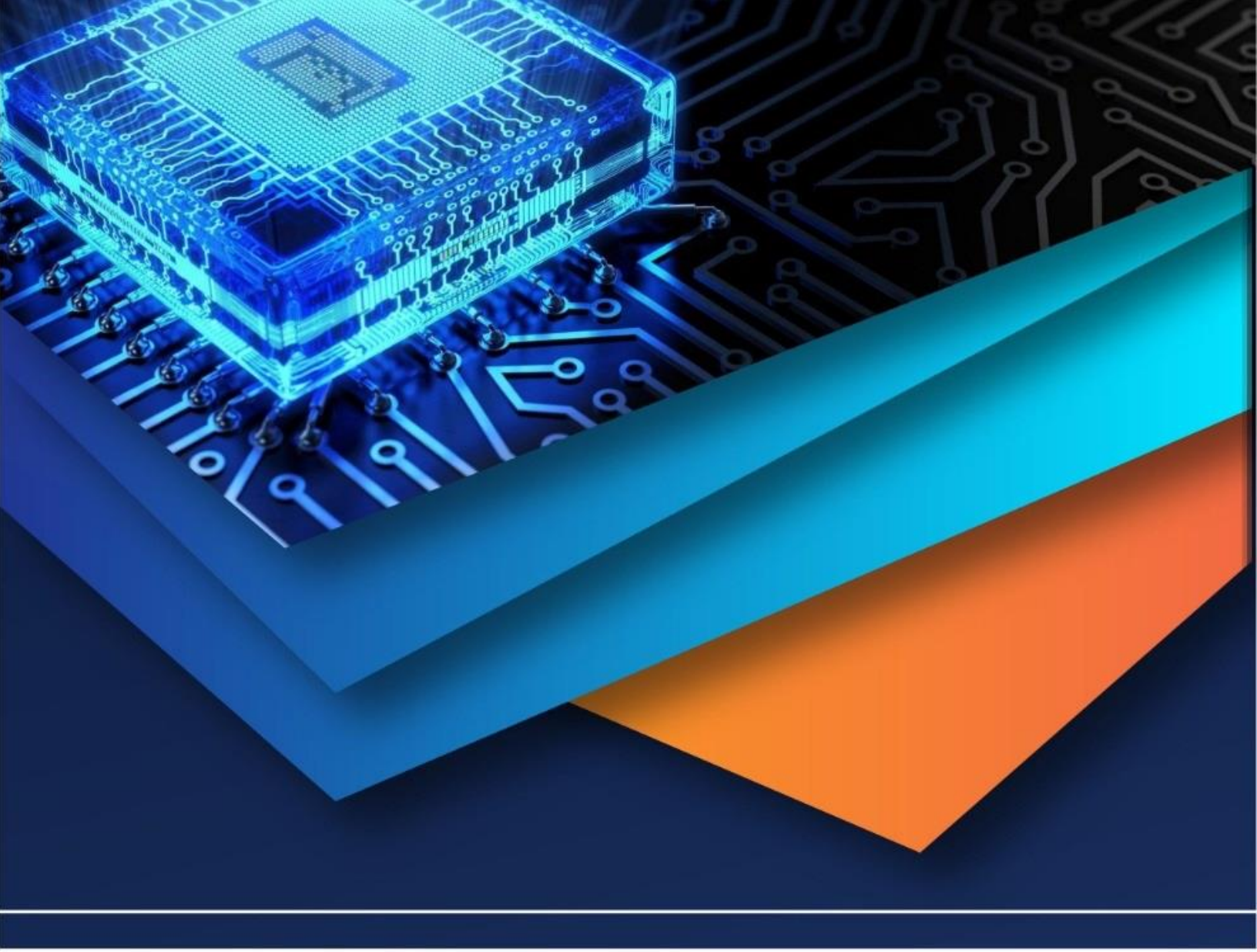
[4] Uday Bhaskar Penta, Dr Panda B S, Dr Sasanko Sekhar Gantayat, “Machine Learning Model For Identifying Phishing Websites”, Journal of Data Acquisition and Processing Volume 38, doi: 10.5281/zenodo.7764722, ISSN: 1004-9037,pp. 2455-2468,2022.

[5] Nikhil K, Dr. Rajesh D S, Dhanush Raghavan, “Phishing Website Detection Using ML”,International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 7, Issue 4, doi : https://doi.org/10.32628/CSEIT217354, ISSN : 2456-3307, pp. 194-198 , July 2021.

[6] Nikita Pawar, Dr. P. A. Tijare, “A Review on Phishing Website Detection Using Machine Learning Approach”, International Journal of Scientific Research in Computer Science, Engineering and Information Technology, Volume 10, Issue 2, ISSN: 2456-3307, pp. 267-272, 9 April 2023.

[7] Hajara musa, Dr. a.y gital, F. U. Zambuk, Abubakar Umar, Aishatu Yahya Umar, Jamilu UsmanWaziri, “A Comparative Analysis of Phishing Website Detection Using Xgboost Algorithm”, Journal of Theoretical and Applied Information Technology, Volume 97, ISSN: 1992-8645, pp. 1434-1443, 15th March 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)