IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Phishing Website Detection using Machine Learning

T. Vyvaswini[1], Mr. P. P Nagaraja Rao[2], B. Kousalya[3], G. Pallavi[4], S. Abdullal[5], P. Siddartha[6]
Associate Professor, Department of Electronics and Communication Engineering[2]
UG Students, Department of Electronics and Communication Engineering[1,3,4,5,6]
Sri Venkatesa Perumal College of Engineering and Technology, Puttur, AP, India

Abstract: *Phishing is a common attack on credulous people by making them to disclose their unique information using counterfeit websites. The objective of phishing website URLs is to purloin the personal information like user name, passwords and online banking transactions. Phishers use the websites which are visually and semantically similar to those real websites. As technology continues to grow, phishing techniques started to progress rapidly and this needs to be prevented by using anti-phishing mechanisms to detect phishing. Machine learning is a powerful tool used to strive against phishing attacks. In this article, we proposed 5 different algorithms in machine learning to analyse the URLs. The accuracy of the Existing method is approximately 94%, and we have implemented it as 95.235% in the Proposed method. Here we used 5 classifiers which are Random Forest Classifier, AdaBoost Classifier, XGBoost Classifier, Support Vector Machine, Gradient Boosting Classifier. Among all these Classifiers, Random Forest Classifier gives the highest accuracy.*

Keywords: AdaBoost, Random forest, XGBoost, performance Analysis, Gradient Boosting and support vector machine

I. INTRODUCTION

Phishing is the most unsafe criminal exercises in cyber space. Since most of the users go online to access the services provided by government and financial institutions, there has been a significant increase in phishing attacks for the past few years. Phishers started to earn money and they are doing this as a successful business. Various methods are used by phishers to attack the vulnerable users such as messaging, VOIP, spoofed link and counterfeit websites. It is very easy to create counterfeit websites, which looks like a genuine website in terms of layout and content. Even, the content of these websites would be identical to their legitimate websites. The reason for creating these websites is to get private data from users like account numbers, login id, passwords of debit and credit card, etc. Moreover, attackers ask security questions to answer to posing as a high level security measure providing to users. When users respond to those questions, they get easily trapped into phishing attacks. Many researches have been going on to prevent phishing attacks by different communities around the world. Phishing attacks can be prevented by detecting the websites and creating awareness to users to identify the phishing websites. Machine learning algorithms have been one of the powerful techniques in detecting phishing websites. In this article, we are using different algorithms in machine learning to detect phising websites.

II. EXISTING ALGORITHM

The Existing system implemented a phising detection system by analysing the URL of the webpage with an accuracy of 94%. The fields such as domain, subdomain, top level domain, protocol, directory, file name, path and query allow creating different URL addresses. These related fields in the phising URLs are generally different from the legitimate ones on websites. The effective features obtained from the URL increase the accuracy of the classification. Additionally, site layout, CSS, content, meta information and other features can also improve accuracy. However, these features will cause an increase in the classification time of the new websites which needed to be classified.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

2.1 Disadvantages:

1. Late process
2. Its more time
3. No accurate result

III. PROPOSED ALGORITHM

Machine Learning is cutting edge and trending for different kinds of diverse application in the society where it can deal with tons of data, refined and revised algorithms, available heavy processing power in terms of GPU. The proposed system aimed to implement python program to extract features from URL. Below are some of the features that we have extracted for detection of phishing URLs.

1. Presence of IP address in URL: If IP address present in URL then the feature is set to 1 else set to 0. Most of the benign sites do not use IP address as an URL to download a webpage. Use of IP address in URL indicates that attacker is trying to steal sensitive information.
2. Presence of @ symbol in URL: If @ symbol present in URL then the feature is set to 1 else set to 0. Phishers add special symbol @ in the URL leads the browser to ignore everything preceeding the "@" symbol and the real address follows after the "@" symbol [6].
3. Number of dots in Host name: Phishing URLs have many dots in URL. For example http://shop.fun.amazon.phishing.com, in this URL phishing.com is an actual domain name, whereas use of number of dots in benign URLs is 3. If the number of dots in URLs is more than 3, then the feature is set to 1 else to 0.
4. Prefix or Suffix separated by (-) to domain: If domain name separated by dash (-) symbol then feature is set to 1 else to 0. The dash symbol is rarely used in legitimate URLs. Phishers add dash symbol (-) to the domain name so that users feel that they are dealing with a legitimate webpage. For example Actual site is http://www.onlineamazon.com but phisher can create another fake website like http://www.online-amazon.com to confuse the innocent users.
5. URL redirection: If "//" present in URL path then feature is set to 1 else to 0. The existence of "//" within the URL path means that the user will be redirected to another website [6].
6. HTTPS token in URL: If HTTPS token present in URL then the feature is set to 1 else to 0. Phishers may add the "HTTPS" token to the domain part of a URL in order to trick users. For example, http://https-www-paypal-it-mpp-home.soft-hair.com [6].
7. Information submission to Email: Phisher might use "mail()" or "mailto:" functions to redirect the user's information to his personal email[6]. If such functions are present in the URL then feature is set to 1 else to 0.
8. URL Shortening Services "TinyURL": TinyURL service allows phisher to hide long phishing URL by making it short. The goal is to redirect user to phishing websites. If the URL is crafted using shortening services (like bit.ly) then feature is set to 1 else 0.

By using above and many other features the accuracy had increased to 95.235%. The classifiers used in proposed system are Random Forest Classifier, AdaBoost Classifier, XGBoost Classifier, Support Vector Machine, Gradient Boosting Classifier. Among all these classifiers, we get highest accuracy for Random Forest Classifier which is approximately equals to 95%.

3.1 Advantages

1. Fast process
2. Less time
3. Accurate result

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

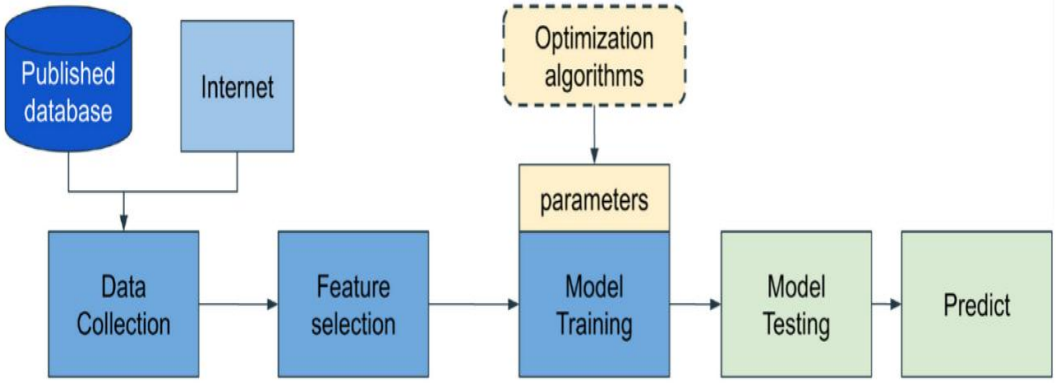

Fig 1: Block Diagram

IV. CLASSIFIERS

4.1 ADABOOST Classifier

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as variance for supervised learning. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones. The AdaBoost algorithm works on the same principle as boosting with a slight difference. First, let us discuss how boosting works. It makes 'n' number of decision trees during the data training period. As the first decision tree/model is made, the incorrectly classified record in the first model is given priority. Only these records are sent as input for the second model. The process goes on until we specify a number of base learners we want to create. Remember, repetition of records is allowed with all boosting techniques.

4.2 XGBOOST

XGBoost stands for "Extreme Gradient Boosting". XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable. It implements Machine Learning algorithms under the Gradient Boosting framework. It provides a parallel tree boosting to solve many data science problems in a fast and accurate way. Boosting is an ensemble learning technique to build a strong classifier from several weak classifiers in series. Boosting algorithms play a crucial role in dealing with bias-variance trade-off. Unlike bagging algorithms, which only controls for high variance in a model, boosting controls both the aspects (bias & variance) and is considered to be more effective XGBoost stands for eXtreme Gradient Boosting. It became popular in the recent days and is dominating applied machine learning and Kaggle competitions for structured data because of its scalability.

4.3 Random Forest Classifier

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems.

A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Random forests reduce the overfitting problem by classifying or averaging the output of individual trees in training processing. Therefore, random forests generally have higher accuracy than decision tree algorithms.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

4.4 Gradient Boosting Classifier

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model. Unlike, Adaboosting algorithm, the base estimator in the gradient boosting algorithm cannot be mentioned by us. The base estimator for the Gradient Boost algorithm is fixed and i.e. Decision Stump. Like, AdaBoost, we can tune the n_estimator of the gradient boosting algorithm. However, if we do not mention the value of n_estimator, the default value of n_estimator for this algorithm is 100. Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier).

4.5 SVM:

A support vector machine (SVM) is a supervised learning algorithm that classifies data points into two sections and predicts new data points belonging to each section. It is suitable for linear binary classification, which has two classes labelled, and the classifier is a hyperplane with N dimensions relevant to the number of features. The core idea of this algorithm is to maximize the distance between the data point and the segmentation hyperplane. For example, there are two classes, phishing and legitimate and a 29-dimension hyperplane when we use the UCI dataset for training the SVM model.

V. RESULTS

Home Page:

Here user view the home page of phishing website prediction web application.

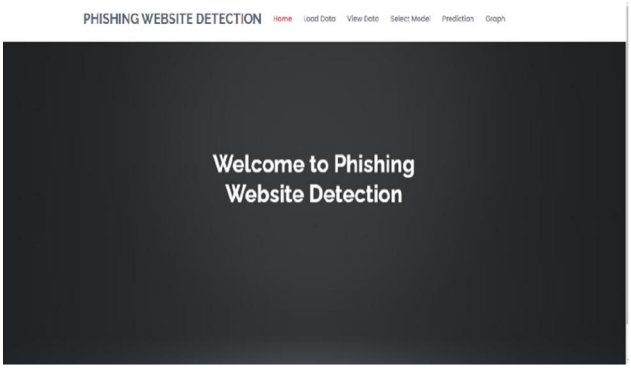

Fig 2: Home Page

Load:

In the load page, users can load the website dataset.

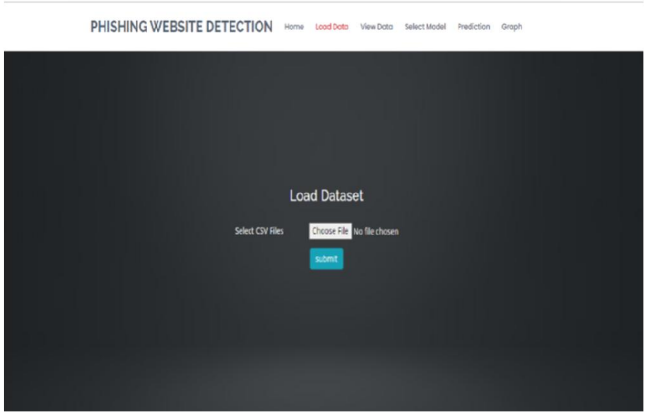

Fig 3: Loading Dataset

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

View:

Here we can see the uploaded data set.

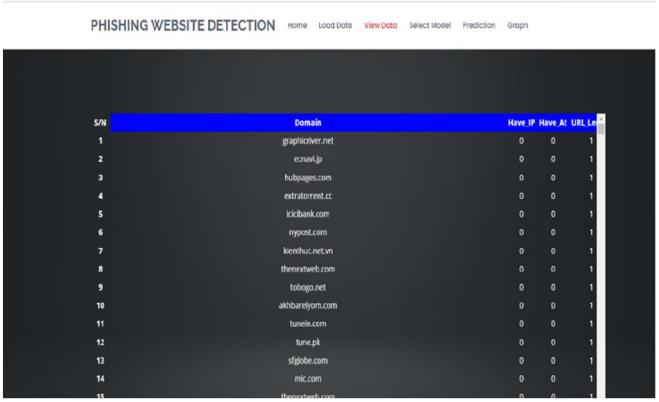

Fig 4: Uploading Dataset

Model:

Here we can train our data using different algorithm.

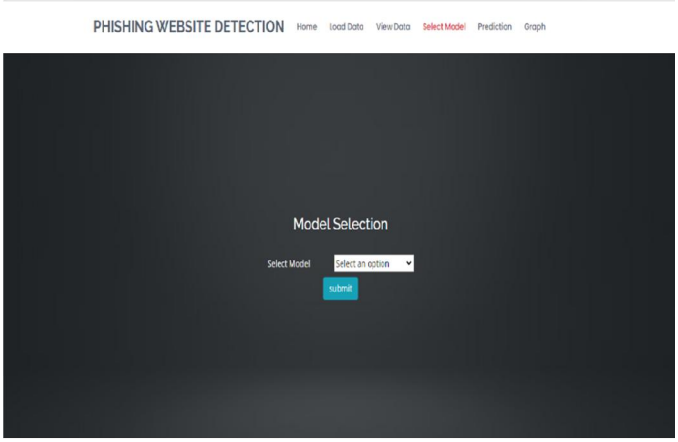

Fig 5: Model

Prediction:

This page show the detection result that whether the website is a phishing website or legitimate.

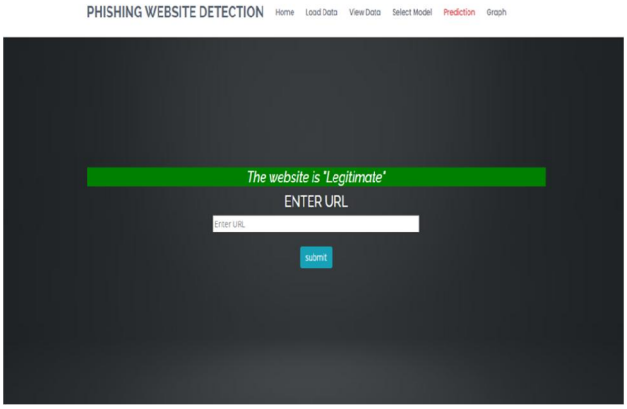

Fig 6: Prediction for Legitimate website

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

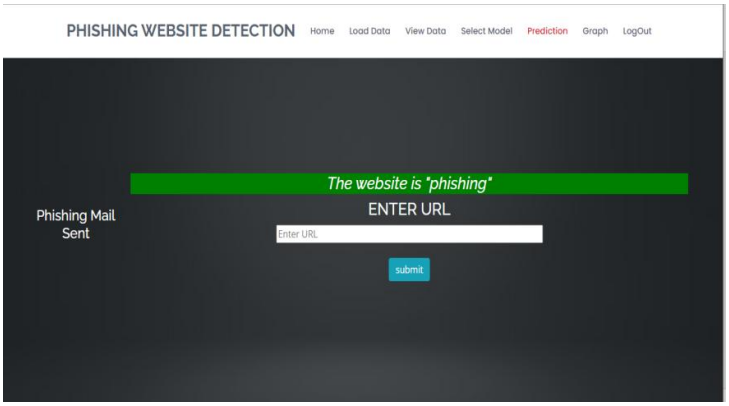

Fig 7: Prediction for Phising website

VI. FLOW CHART OF CODING PROCESS

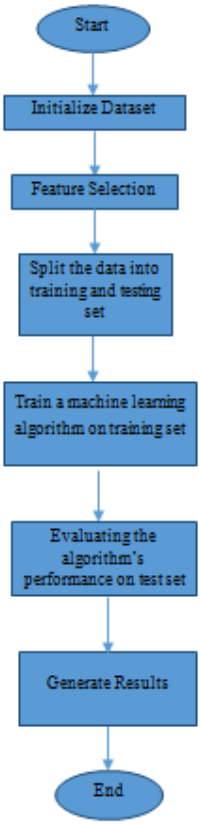

Fig 8: Flow chart of coding process

VII. MODULES

User

View Home page

Here user view the home page of the phishing website prediction web application.

View Upload page

In the about page, users can learn more about the phishing prediction.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

Input Model
The user must provide input values for the certain fields in order to get results.
View Results
User view's the generated results from the model.
View score
Here user have ability to view the score in %

7.1 System
Working on dataset
System checks for data whether it is available or not and load the data in csv files.
Pre-processing
Data need to be pre-processed according the models it helps to increase the accuracy of the model and better information about the data.
Training the data
After pre-processing the data will split into two parts as train and test data before training with the given algorithms. The dataset that is available in [18] and [20] are used in training process.
Model Building
To create a model that predicts the personality with better accuracy, this module will help user.
Generated Score
Here user view the score in percentage (%).
Generate Results
We train the machine learning algorithm and calculate the personality prediction.
The design of output is the most important task of any system. During output design, developers identify the type of outputs needed, and consider the necessary output controls and prototype report layouts. System checks for data whether it is available or not and load the data in csv files. Data need to be pre-processed according the models it helps to increase the accuracy of the model and better information about the data. After pre-processing the data will split into two parts as train and test data before training with the given algorithms. To create a model that predicts the personality with better accuracy, this module will help user.

7.2 Test Results of Classifiers:

Serial No.	Classifier	Existing Accuracy	Proposed Accuracy
1	XGBoost	81.69	82.9807
2	Random Forest	94.59	95.2354
3	SVM	70.2	79.0384
4	AdaBoost	-	79.1025
5	Gradient Boosting	-	80.9935

Table 1: Test Results of Classifiers

VIII. CONCLUSION

Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This article deals with machine learning technology for detection of phishing URLs by extracting and analysing various features of legitimate and phishing URLs. Phishing attacks can be prevented by detecting the websites and creating awareness to users to identify the phishing websites. Machine learning algorithms have been one of the powerful techniques in detecting phishing websites. The Machine Learning algorithms are XGBoost, AdaBoost, SVM, Gradient Boosting and Random Forest classifiers. The main drawback in existing method is that it gives 94% accuracy. In our article, we improved the accuracy to 95%. This article gives highest accuracy around 95% using Random Forest Classifier. The main advantage of this article is, it gives accurate results.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

IX. FUTURE SCOPE

There are quite a few things that can be polished or be added in the future work. We have opted to use data mining classifiers in this article namely the XGBoost, AdaBoost, SVM, Gradient Boosting and Random Forest classifiers. There are more classifiers such as the Bayesian network classifier, Neural Network classifier and C4.5 classifier. Such classifiers were not used in this experiment but may be used in the future to provide more data for comparison.

Despite there are several ways to carry out these attacks, unfortunately the current phishing detection techniques cover some attack vectors like email and fake websites. Therefore, building a specific limited scope detection system will not provide complete protection from the wide phishing attack vectors. This article develops detection system with a wide protection scope using URL features only which is relying on the fact that users directly deal with URLs to surf the internet and provides a good approach to detect malicious URLs as proved by previous studies.

REFERENCES

[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]. [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]. Mohammad R., Thabtah F. McCluskey L. (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016.

[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. 16.

[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]. Sharma, Ushamary and Ghisingh, Seema and Ramdinmawii, Esther, "A Study on the Cyber - Crime and Cyber Criminals: A Global Problem," International Journal of Web Technology, vol 03, pp. 172-179, June 2014.

[12]. Andrewa, "Cybercrime", http://en.wikipedia.org/wiki/Computer_crime, October 15, 2003.

[13]. Vayansky, I. and Kumar, S., "Phishing – challenges and solutions.", Computer Fraud & Security, vol 2018, no. 1, pp. 15-20, January 2018.

[14]. Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, "Phishing Detection Using Machine Learning Techniques," unpublished.

[15]. Gokula Chandar ,Leeban Moses M; T. Perarasi M; Rajkumar; "Joint Energy and QoS-Aware Cross-layer Uplink resource allocation for M2M data aggregation over LTE-A Networks", IEEE explore, doi:10.1109/ICAIS53314.2022.9742763.

[16]. Mustafa Alper Akkaş, Radosveta Sokullu, "An IoT-based greenhouse monitoring system with Micaz motes", https://doi.org/10.1016/j.procs.2017.08.300.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 1, February 2023
Impact Factor: 7.301

[17]. P. V. Vimal and K. S. Shivaprakasha, "IOT based greenhouse environment monitoring and controlling system using Arduino platform," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, 2017, pp. 1514-1519.

[18]. Dhuddu Haripriya, Venkatakiran S, Gokulachandar A, "UWB-Mimo antenna of high isolation two elements with wlan single band-notched behavior using roger material", Vol 62, Part 4, 2022, Pg 1717-1721, https://doi.org/10.1016/j.matpr.2021.12.203.

[19]. Gokula Chandar A, Vijayabhasker R., and Palaniswami S, "MAMRN – MIMO antenna magnetic field", Journal of Electrical Engineering, vol.19, 2019.

[20]. Rukkumani V , Moorthy V, Karthik M , Gokulachandar A, Saravanakumar M, Ananthi P, "Depiction of Structural Properties of Chromium Doped SnO2 Nano Particles for sram Cell Applications", Journal of Materials Today:

BIBLIOGRAPHY

Thirunagari Vyvaswini, UG Student,
Dept of ECE, SVPCET
Area of Interest- Machine Learning.

Boligarla Kousalya, UG Student,
Dept of ECE, SVPCET
Area of Interest- Machine Learning.

Gummalla Pallavi, UG Student,
Dept of ECE, SVPCET
Area of Interest- Machine Learning.

S. Abdullal, UG Student,
Dept of ECE, SVPCET
Area of Interest- Machine Learning.

Pangala Siddartha, UG Student,
Dept of ECE, SVPCET
Area of Interest-
Machine Learning.