PHISHING ATTACK DETECTION USING MACHINE LEARNING

A PROJECT REPORT

(Project Report Phase – I)

Submitted by

NITISH KUMAR R - 212219220036
DAMODARRAM K - 212219220005
GOPI G - 212219220009

in partial fulfillment for the award of the degree

of

BACHELOR OF TECHNOLOGY

IN

INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION TECHNOLOGY

SAVEETHA ENGINEERING COLLEGE, THANDALAM.

ANNA UNIVERSITY: CHENNAI 600 025

DEC 2022

BONAFIDE CERTIFICATE

Certified that this project report "**PHISHING ATTACK DETECTION USING MACHINE LEARNING**" is the Bonafide work of **NITISH KUMAR R (212219220036), DAMODARRAM K (212219220005), GOPI G (212219220009),** who carried out the project work under my supervision.

SIGNATURE
Dr. K. Suresh Kumar, M.Tech, Ph.D
HEAD OF THE DEPARTMENT
ASSOCIATE PROFESSOR
Department of Information Technology,
Saveetha Engineering College,
Saveetha Nagar, Thandalam,
Chennai-602105

SIGNATURE
Dr. G. Nalini Priya, M.E, Ph.D
SUPERVISOR
PROFESSOR
Department of Information Technology,
Saveetha Engineering College,
Saveetha Nagar, Thandalam,
Chennai-602105

Submitted for VIVA-VOCE held on **……**

INTERNAL EXAMINER **EXTERNAL EXAMINER**

ACKNOWLEDGEMENT

I wish to express my gratitude to our Founder President **Dr. N.M. Veeraiyan**, Director **Dr. S. Rajesh,** Saveetha Engineering College, for their guidance and blessings.

I am very grateful to our Principal **Dr. N. Duraipandian M.E., Ph.D.,** for providing me with an environment to complete my project successfully.

I am indebted to our **Head of the Department, Dr. K. Suresh Kumar, M.Tech., Ph.D., for** his support during the entire course of this project work.

I am indebted to our **supervisor Project, Dr G. NALINI PRIYA , M.E., Ph.D Professor,** Department of Information Technology, Saveetha Engineering College, for assisting me in the completion of my project with her exemplary Guidance and for his/her support during the entire course of this project.

My heartfelt thanks to the **Project Coordinator, Dr. K. Suresh Kumar, M.Tech., Ph.D., Associate Professor,** Department of Information Technology, Saveetha Engineering College, for unstinted support throughout this project.

I also thank all the staff members of our department for their help in making this project successful.

ABSTRACT

Phishing sites which expects to take the victims confidential data by diverting them to surf a fake website page that resembles a honest to goodness one is another type of criminal acts through the internet and its one of the especially concerns toward numerous areas including e-managing an account and retailing. Phishing site detection is truly an unpredictable and element issue including numerous components and criteria that are not stable. On account of the last and in addition ambiguities in arranging sites because of the intelligent procedures programmers are utilizing, some keen proactive strategies can be helpful and powerful tools can be utilized, for example, fuzzy, neural system and data mining methods can be a successful mechanism in distinguishing phishing sites. We applied Random Forest (RF), one of the different types of machine learning based algorithms used for detection of Phishing websites. Finally we measured and compared the performance of the classifier in terms of accuracy.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Phishing is a type of extensive fraud that happens when a malicious website act like a real one keeping in mind that the end goal to obtain touchy data, for example, passwords, account points of interest, or MasterCard numbers. In spite of the fact that there are a few contrary to phishing programming and methods for distinguishing potential phishing endeavours in messages and identifying phishing substance on sites, phishes think of new and half breed strategies to go around the accessible programming and systems. Phishing is a trickery system that uses a blend of social designing what's more, innovation to assemble delicate and individual data, for example, passwords and charge card subtle elements by taking on the appearance of a dependable individual or business in an electronic correspondence. Phishing makes utilization of spoof messages that are made to look valid and implied to be originating from honest to goodness sources like money related foundations, ecommerce destinations and so forth, to draw clients to visit fake sites through joins gave in the phishing email. The misleading sites are intended to emulate the look of a genuine organization site page. The employing so as to phishing invader's trap clients diverse social building strategies, for example, debilitating to suspend client accounts on the off chance that they don't finish the account upgrade process, give other data to approve their records or a few different motivations to get the clients to visit their satirize page. Supervised learning (Classification Technique) accommodates a vastly improved precision while unsupervised learning accommodates a quick and dependable way to deal with infer information from a dataset. That's why we used supervised learning in our work.

1.1 OBJECTIVE

1. Understanding phishing domain (or Fraudulent Domain) characteristics, its distinguishing features from legitimate domains
2. Why it is so important to detect this domain and how they can be detected using machine learning and natural language processing techniques
3. Reviewing the state-of-the-art machine learning techniques for malicious URL detection in literature
4. Understanding the newly emerging concept of Malicious URL Detection as a service and the principles to be used while designing such a system.

To distinguish the phishing websites from the legitimate websites and ensure secure transactions to users

1.2 SCOPE OF THE PROJECT

1. The system should be useful in many e-commercial websites for maintaining the security and reliability of customers and people online
2. The system should be useful in preventing online frauds leading to leakage of important and private user data
3. The scope of using Machine Language over other Traditional Detecting Methods

CHAPTER 2

LITERATURE REVIEW

Mohammad Nazmul Alam; Dhiman Sarma; Farzana Firoz Lima; Ishita Saha; Rubaiath-E- Ulfath; Sohrab Hossain, 2020, "Phishing Attacks Detection using Machine Learning Approach", IEEE Smart Systems and Inventive Technology,Volume:50 Issue:6

Evolving digital transformation has exacerbated cybersecurity threats globally. Digitization expands the doors wider to cybercriminals. Initially cyberthreats approach in the form of phishing to steal the confidential user credentials. Usually, Hackers will influence the users through phishing in order to gain access to the organizatlou's digital assets and networks. With security breaches, cybercriminals execute ransomware attack, get unauthorized access, and shut down systems and even demand a ransom for releasing the access. Anti-phishing software and techniques are circumvented by the phishers for dodging tactics. Though threat inte'ligence and behavioural analytics systems support organizations to spot the unusual traffic patterns, still the best practice to prevent phishing attacks is defended in depth. In this perspective, the proposed research work has developed a model to detect the phishing attacks using machine learning (ML) algorithms like random forest (RF) and decision tree (DT). A standard legitimate dataset of phishing attacks from Kaggle was aided for ML processing. To analyze the attributes of the dataset, the proposed model has used feature selection algorithms like principal component analysis (PCA). Finally, a maximum accuracy of 97% was achieved through the random forest algorithm.

JianMao ,JingdongBian ,WenqianTian ,ShishiZhu ,TaoWei ,AiliLi ,ZhenkaiLiang/ 2018, 'Detecting Phishing Websites via Aggregation Analysis of Page Layouts', International Research Journal of Engineering and Technology (IRJET), Volume: 08 Issue: 11.

In this paper, we aim to improve phishing detection techniques using machine learning techniques. In particular, we propose a learning-based aggregation analysis mechanism to decide page layout similarity, which is used to detect phishing pages. Our experiment results shows that our approach is accurate and effective in detecting phishing pages. Of the 95,321 phishing domains, we identified 27,253 domain names that we believe were registered maliciously, by phishers,"."This is an all-time high, and much higher than the

22,629 we identified in 1H2014. Most of these registrations were made by Chinese phishers. The other 68,303 domains were almost all hacked or compromised on vulnerable Web hosting."[' Global Phishing Survey: Trends and Domain Name Use in 2H2014']

Atharva Deshpande , Omkar Pedamkar , Nachiket Chaudhary , Dr. Swapna Borde/ 2021," Detection of Phishing Websites using Machine Learning", Volume: 07 Issue: 06 | June 2018

This paper surveys the features used for detection and detection techniques using machine learning.Phishing is popular among attackers, since it is easier to trick someone into clicking a malicious link which seems legitimate than trying to break through a computers defense systems. The malicious links within the body of the message are designed to make it appear that they go to the spoofed organization using that organizations logos and other legitimate contents.We identified 27,253 domain names that we believe were registered maliciously, by phishers. This is an all-time high, and much higher than the 22,629 we identified in 1H2014. Most of these registrations were made by Chinese phishers. The other 68,303 domains were almost all hacked or compromised on the vulnerable Web hosting.

Ishant Tyagi; Jatin Shad; Shubham Sharma; Siddharth Gaur; Gagandeep Kaur/ 2018et al," A Novel Machine Learning Approach to Detect Phishing Websites, February 2018, IEEE Access, Volume: 8, ISSN: 2169-3536

This paper is focused on various Machine Learning algorithms aimed at predicting whether a website is phishing or legitimate. Machine learning solutions are able to detect zero hour phishing attacks and they are better at handling new types of phishing attacks, so they are preferred. In our implementation, we managed an accuracy of 98.4% in prediction a website to be phishing or legitimate. Phishers continued to attack Apple, PayPal, and Taobao.com heavily. Each of these three e-commerce giants suffered over 20,000 phishing attacks against their respective services and brands. Together, these top three were the targets of nearly 54 percent of the world's phishing attacks. The next seven brands were targeted for a combined 23 percent of all phishing attacks — meaning the top 10 targets accounted for over three quarters of all phishing attacks observed worldwide. The number of times that the targets were attacked follows a long tail. Half of the targets were attacked four or fewer times during the six-month period.

Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi/ 2020," Phishing Detection Using Machine Learning TechniquesSands, Singapore, 10-12th December 2019

In this paper, we propose a novel technique to identify phishing websites effortlessly on the client side by proposing a novel browser architecture. In this system, we use the rule of extraction framework to extract the properties or features of a website using the URL only. This list consists of 30 different properties of a URL, which will later be used by the Random Forest Classification machine learning model to detect the authenticity of the website. These show criminals seeking the credentials of consumers in places where consumers may least expect it. Phishers target wide-ranging targets for several reasons. One is to perform credit card theft, and hitting new targets may lull consumers into a false sense of security. The phishers can also monetize stolen data through reshipping fraud, a tactic that remains popular. Phishers also steal usernames and passwords from one site in order to try those credential on other sites. Many consumers re-use usernames and passwords, and this poor habit can be costly. If a site is getting phished for the first time, it may have been targeted by a more sophisticated phisher, who had the skill to design a new phishing template

CHAPTER 3

SYSTEM ANALYSIS

4.1 EXISTING SYSTEM

- This paper approaches a framework to extract features flexible and simple with new strategies. Data is collected from Phish Tank and legitimate URLs from Google.
- To obtain the text properties C# programming and R programming were used.
- 133 features were obtained from the dataset and third party service providers. CFS subset based and Consistency subset based feature selection methods used for feature selection and analyzed with WEKA tool.
- Naïve Bayes and Sequential Minimal Optimization (SMO) algorithms were compared for performance evaluation and SMO is preferred by the author for phishing detection than NB.

3.2 PROPOSED SYSTEM

- In this work we aim to review various machine learning methods used for this purpose, along with datasets and URL features used to train the machine learning models.
- A URL based phishing attack is carried out by sending malicious links, that seems legitimate to the users, and tricking them into clicking on it. In phishing detection, an incoming URL is identified as phishing or not by analysing the different features of the URL and is classified accordingly. Different machine learning algorithms are trained on various datasets of URL features to classify a given URL as phishing or legitimate.

- We achieved 97.14% accuracy for Random Forest algorithm with the lowest false negative rate. The paper concluded that accuracy increases when more data is used for training

CHAPTER 4

SYSTEM DESIGN

4.2 SYSTEM ARCHITECTURE

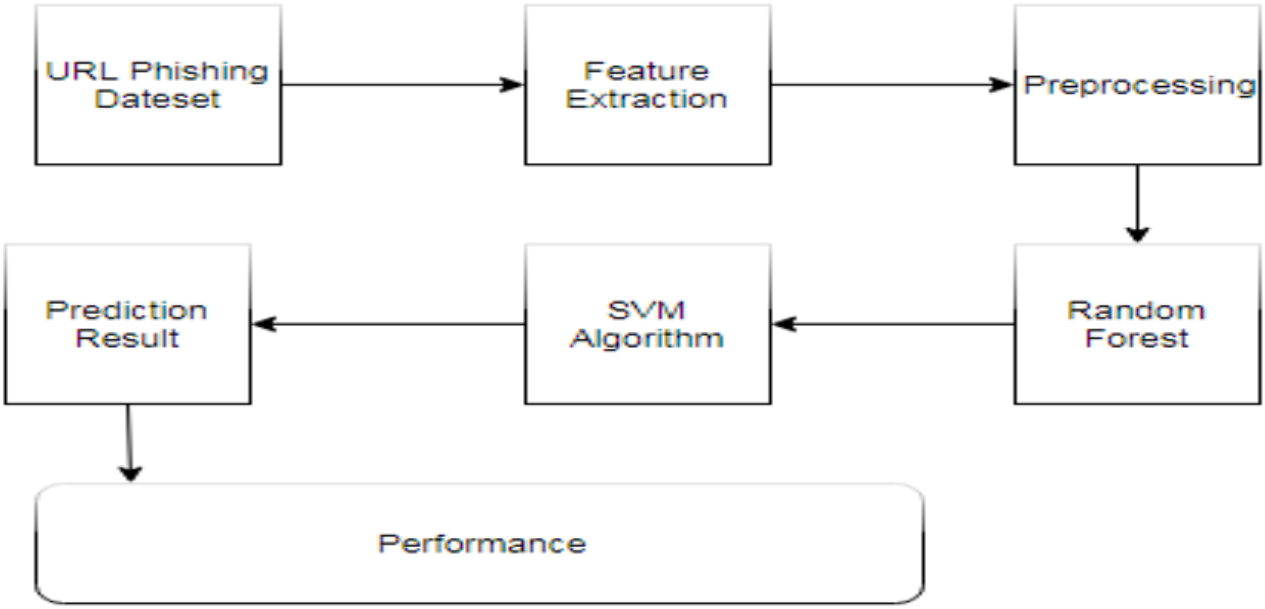

Fig 4.1

4.2 SYSTEM REQUIREMENTS

4.2.1 Hardware Requirements

- System : Intel Pentium IV 2.80 GHz.
- Monitor : LED.
- Mouse : Logitech.
- Ram : 4.00 GB or above 4.00 GB
- Hard Disk : 250 GB

4.2.2 Software Requirements:

- Operating system : Windows 7, Ubuntu
- Language : Python 3

4.3 MODULES

4.3.1 LIST OF MODULES

- ✓ Detection Technique
- ✓ Phishing Websites Features
- ✓ Data Set

4.3.2 MODULE 1: DETECTION TECHNIQUE

Detection of phishing websites has received a lot of attention recently due to their impact on users' security. Therefore, many techniques have been developed to detect phishing websites varying from communication-oriented techniques, such as authentication protocols, blacklisting, and white-

listing, to content-based filtering techniques. The blacklisting and white-listing techniques have not proven though to be sufficiently efficient when used in different domains, and thus they are not commonly used. Meanwhile, the content-based phishing filters have been widely used and have proven to be of high efficiency. In light of this, researches have focused on content-based mechanism and on developing machine learning and data mining techniques based on the header and body of emails.

4.3.3 MODULE 2: PHISHING WEBSITES FEATURES

One of the challenges faced by us was the unavailability of reliable training datasets. In fact, this challenge faces any researcher in the field. However, although plenty of articles about predicting phishing websites using data mining techniques have been disseminated these days, no reliable training dataset has been published publically, maybe because there is no agreement in literature on the definitive features that characterize phishing websites, hence it is difficult to shape a dataset that covers all possible features. In this article, we shed light on the important features that have proved to be sound and effective in predicting phishing websites. In addition, we proposed some new features, experimentally assign new rules to some well-known features and update some other features.

4.3.4 MODULE 3: DATA SET

The dataset is collected from the kaggle website which is an open source. It consists of 16 features and more than 1000 records.

4.4 SYSTEM IMPLEMETATION

4.4.1 INPUT DESIGN AND OUTPUT DESIGN

INPUT DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.
- Methods for preparing input validations and steps to follow when error occur.

OBJECTIVES

1. Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

2. It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

3. When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

OUTPUT DESIGN

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

2. Select methods for presenting information.

3. Create document, report, or other formats that contain information produced by the system.

The output form of an information system should accomplish one or more of the following objectives.

- Convey information about past activities, current status or projections of the
- Future.
- Signal important events, opportunities, problems, or warnings.
- Trigger an action.
- Confirm an action.

4.4.2 DATA FLOW DIAGRAM:

1. The DFD is also called as bubble chart. It is a simple graphical formalism that can be used to represent a system in terms of input data to the system, various processing carried out on this data, and the output data is generated by this system.
2. The data flow diagram (DFD) is one of the most important modeling tools. It is used to model the system components. These components are the system process, the data used by the process, an external entity that interacts with the system and the information flows in the system.

3. DFD shows how the information moves through the system and how it is modified by a series of transformations. It is a graphical technique that depicts information flow and the transformations that are applied as data moves from input to output.
4. DFD is also known as bubble chart. A DFD may be used to represent a system at any level of abstraction. DFD may be partitioned into levels that represent increasing information flow and functional detail.

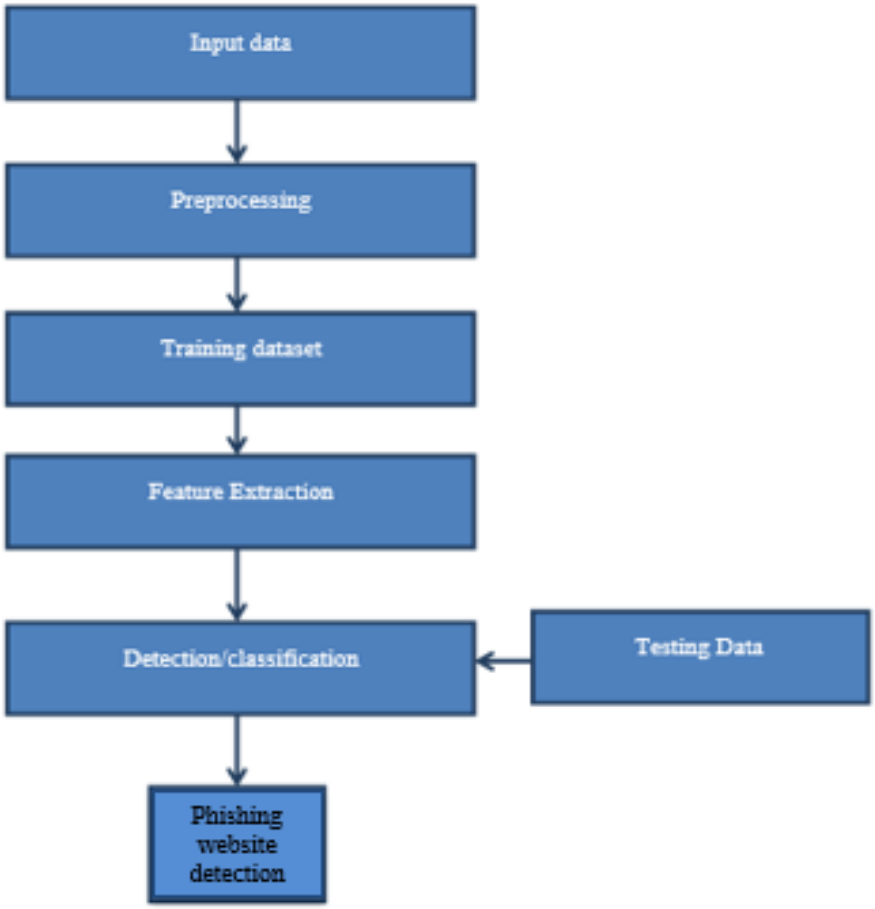

Fig 4.2

4.4.3 UML DIAGRAMS

UML stands for Unified Modeling Language. UML is a standardized general-purpose modeling language in the field of object-oriented software engineering. The standard is managed, and was created by, the Object Management Group.

The goal is for UML to become a common language for creating models of object oriented computer software. In its current form UML is comprised of two major components: a Meta-model and a notation. In the future, some form of method or process may also be added to; or associated with, UML.

The Unified Modeling Language is a standard language for specifying, Visualization, Constructing and documenting the artifacts of software system, as well as for business modeling and other non-software systems.

The UML represents a collection of best engineering practices that have proven successful in the modeling of large and complex systems.

The UML is a very important part of developing objects oriented software and the software development process. The UML uses mostly graphical notations to express the design of software projects.

GOALS:

The Primary goals in the design of the UML are as follows:

1. Provide users a ready-to-use, expressive visual modeling Language so that they can develop and exchange meaningful models.
2. Provide extendibility and specialization mechanisms to extend the core concepts.
3. Be independent of particular programming languages and development process.

4. Provide a formal basis for understanding the modeling language.
5. Encourage the growth of OO tools market.
6. Support higher level development concepts such as collaborations, frameworks, patterns and components.
7. Integrate best practices.

4.4.4 USE CASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

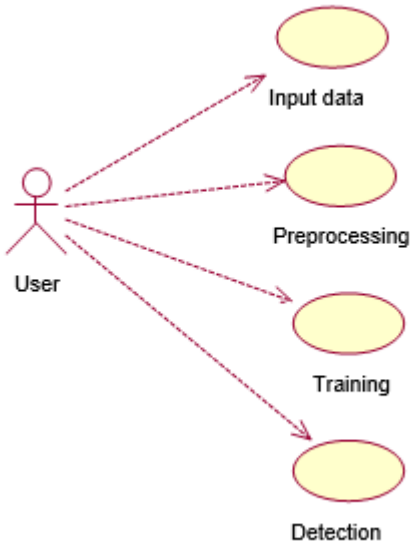

Fig 4.3

CLASS DIAGRAM:

In software engineering, a class diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among the classes. It explains which class contains information.

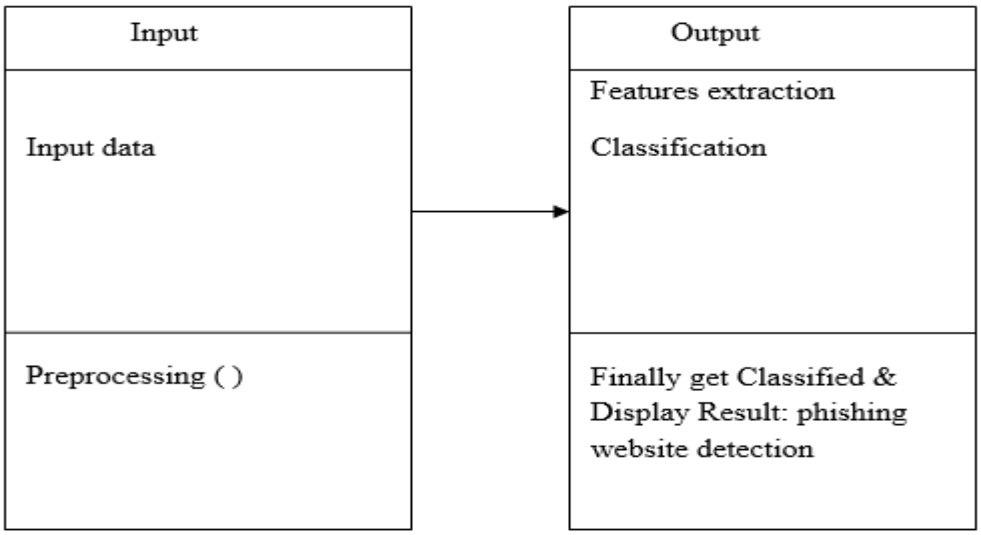

Fig 4.4

4.4.5 SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

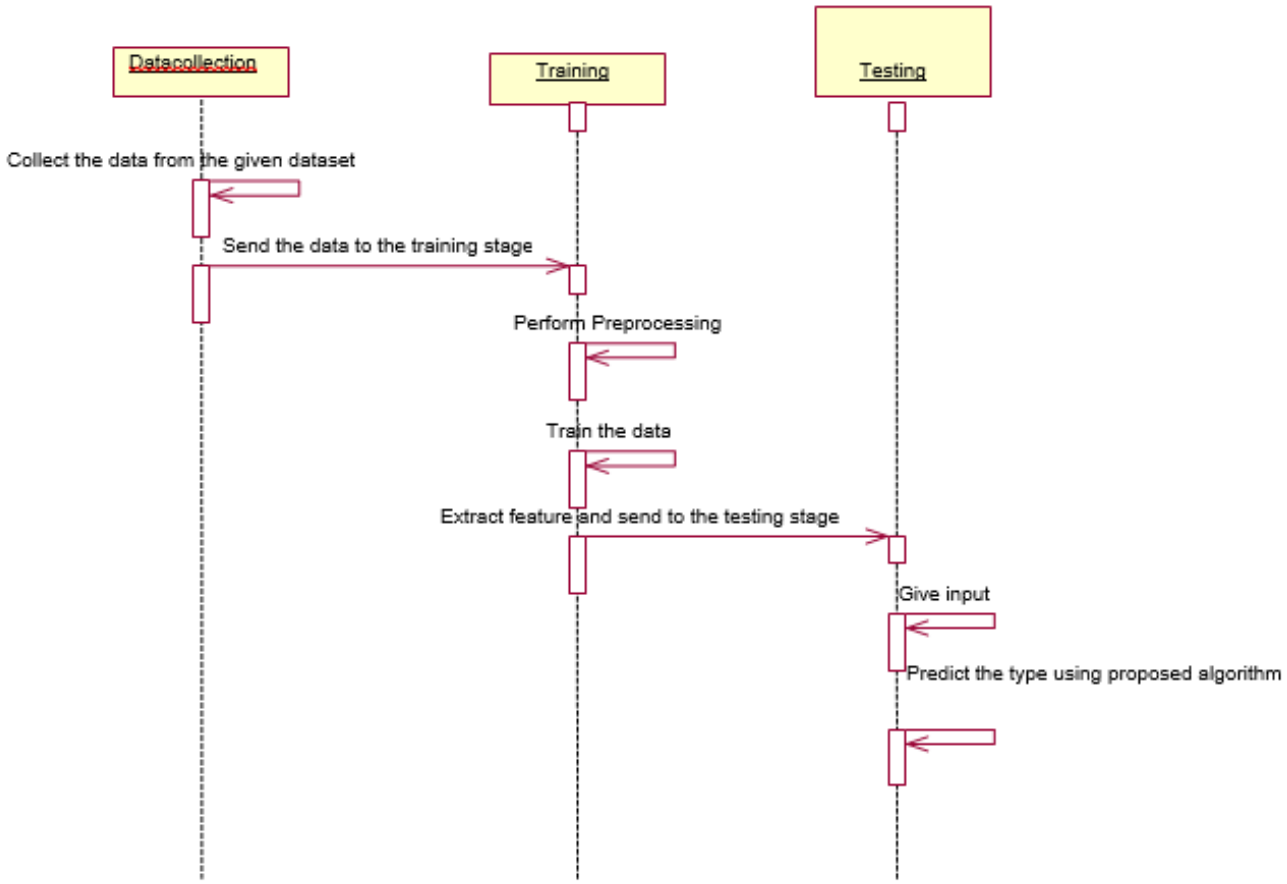

Fig 4.5

4.4.6 ACTIVITY DIAGRAM:

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. An activity diagram shows the overall flow of control.

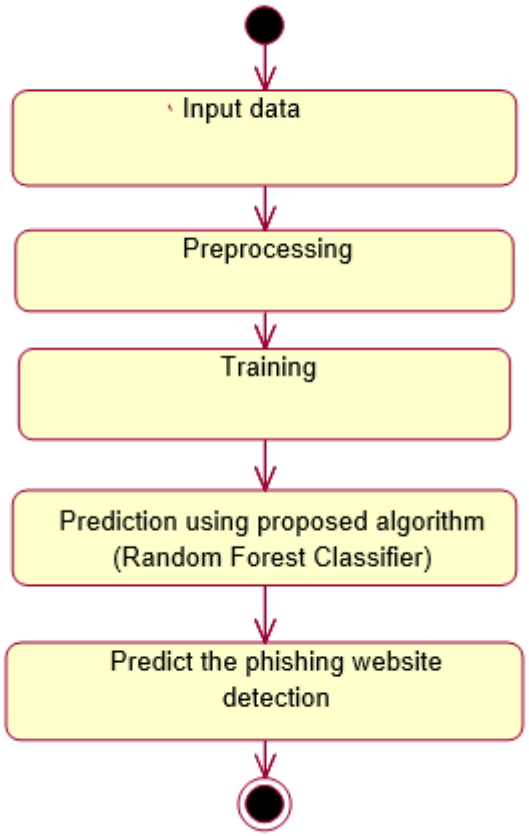

Fig 4.6

CHAPTER 5

5.1 CONCLUSION

Phishing is a cyber crime procedure utilizing both social building and specialized deception to take individual sensitive data. Besides, Phishing is considered as another extensive type of fraud. Experimentations against recent dependable phishing data sets utilizing different classification algorithm have been performed which received different learning methods. The base of the experiments is accuracy measure.

The aim of this research work is to predict whether a given URL is phishing website or not. It turns out in the given experiment that Random forest based classifiers are the best classifier with great classification accuracy of 75.47% for the given dataset of phishing site. As a future work we might use this model to other Phishing dataset with larger size then now and then testing the performance of those classification algorithm's in terms of classification accuracy.

5.2 FUTURE WORK

Future work will aim to develop a system that can learn by itself about new types of phishing attacks by adding a more enhanced feature to the detection process. The scope of this approach not only helps in adding more enhanced features but also updating the existing features to improve its importance level to make detection more efficient and reduce the false positive rate to a large extent. Another further work should include deploying this approach into a web extension to make the detection more robust to the user.

REFERENCES

[1] Samuel Marchal, Jérôme François, Radu State, and Thomas Engel, "PhishStorm: Detecting Phishing With Streaming Analytics," IEEE Transactions on Network and Service Management, vol. 11 , issue: 4 , pp. 458-471, December 2014.

[2] Mohammed Nazim Feroz,Susan Mengel, "Phishing URL Detection Using URL Ranking," IEEE International Congress on Big Data, July 2015.

[3] Mahdieh Zabihimayvan, Derek Doran, "Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection," International Conference on Fuzzy Systems (FUZZ-IEEE), New Orleans, LA, USA, June 2019.

[4] Moitrayee Chatterjee,Akbar-Siami Namin, "Detecting Phishing Websites through Deep Reinforcement Learning," IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), July 2019.

[5] Chun-Ying Huang,Shang-Pin Ma,Wei-Lin Yeh,Chia-Yi Lin,ChienTsung Liu, "Mitigate web phishing using site signatures," TENCON 2010-2010 IEEE Region 10 Conference, January 2011.

[6] Aaron Blum,Brad Wardman,Thamar Solorio,Gary Warner, "Lexical feature based phishing URL detection using online learning," 3rd ACM workshop on Artificial intelligence and security, Chicago, Illinois, USA, pp. 54-60, August 2010.

[7] Mohammed Al-Janabi,Ed de Quincey,Peter Andras, "Using supervised machine learning algorithms to detect suspicious URLs in online social networks," IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, Sydney, Australia, pp. 1104-1111, July 2010.

[8] Erzhou Zhu,Yuyang Chen,Chengcheng Ye,Xuejun Li,Feng Liu, “OFSNN:An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network,” IEEE Access(Volume:7), pp. 73271-73284, June 2019.

[9] Ankesh Anand,Kshitij Gorde,Joel Ruben Antony Moniz,Noseong Park,Tanmoy Chakraborty,Bei-Tseng Chu, “Phishing URL Detection with Oversampling based on Text Generative Adversarial Networks,” IEEE International Conference on Big Data (Big Data), December 2018.

[10] Justin Ma,Lawrence K. Saul,Stefan Savage,Geoffrey M. Voelker, “Learning to detect malicious URLs,” ACM Transactions on Intelligent Systems and Technology (TIST) archive Volume 2 Issue 3, April 2011.