

How to Detect Phishing Website Using Three- Model Ensemble Classification

كيفية اكتشاف موقع التصيد الاحتيالي باستخدام تصنيف
المجموعة ثلاثية النماذج

Prepared By

Yussra M. AL-Shareef

Supervisor

Dr. Hesham Abusaimh

A Thesis Submitted in Partial Fulfilment of the Requirements
of the Master Degree in Computer Science

Computer Science Department
Faculty of Information Technology
Middle East University

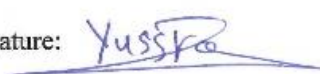
June, 2020

Authorization

I, Yusra M. AL-Shareef, Authorized the Middle East University to provide hard copies or soft copies of my thesis to libraries, institutions, or individuals upon their request.


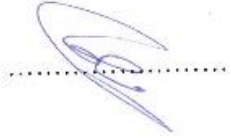
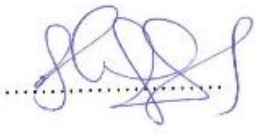
Name: Yusra M. AL-Shareef

Date: 13/06/2020

Signature: 

Thesis Committee Decision

This is to certify that the thesis entitle "**How To Detect Phishing Website Using Three-Model Ensemble Classification**" was successfully defended and provide on 03/06/2020.

| Examination Committee Members | Signature |
|--|---|
| <p>Dr. Hesham Abusaimh (Supervisor/chairman)</p> <p>Associate professor, Department of Computer Science</p> <p>Dean of International Programmes</p> <p>Dean of Graduate Studies and Scientific Research</p> <p>Middle East University(MEU)</p> |  |
| <p>Dr. Mudahfar Al-Jarrah (Internal Examiner)</p> <p>Associate Professor, Department of Computer Science</p> <p>Middle East University(MEU)</p> |  |
| <p>Dr. Mohammad Shkoukani (External Examiner)</p> <p>Department of Computer Science</p> <p>Applied Science University (ASU)</p> |  |

Acknowledgment

Many thanks are submitted first and foremost to Allah who gave me the strength and ability to complete this study.

knowledge, confidence, patience to pass this Master thesis successfully. Also, I owe a great gratitude for those who inspired me throughout this venture to express my thank to my thesis advisor. Dr. Hesham Abusaimh for the complete guidance throughout the thesis stages, and for the critical assistance in designing and proceeding the methodology of my research. I would also like to express my appreciation to the Middle East University and Department of Computer Science where I spent great times.

Finally, I thank all those, who have helped me directly or indirectly in the successful completion of my research work.

Yussra M. AL-Shareef

The Researcher

Dedication

Every challenging work needs self-efforts as well as the guidance of older especially those who were very close to our heart. This study dedicated to my whole family and friends; **My Mother**, no words can describe what you have done for me, thank you for your endless love. My sweetest brothers, who are one part of my life. I would also like to dedicate this thesis to the spirit of **My Father's** who supported me in every step of my life and encouraging me believed in myself. It is hard to find a word to express my gratitude and thanks, to each of the above, I extend my deepest appreciation.

Yussra M. AL-Shareef

Table of Contents

| | |
|---|-----------|
| Cover Page..... | I |
| Authorization..... | II |
| Thesis Committee Decision..... | III |
| Table of Contents | VII |
| List of Tables | III |
| List of Figures..... | IV |
| List of Abbreviations | V |
| Abstract..... | VI |
| المُلخَص..... | VII |
| Chapter One: Introduction | 1 |
| 1.1 Introduction..... | 2 |
| 1.2 Types of phishing Attacks | 4 |
| 1.2.1 Clone Phishing Attack | 4 |
| 1.2.2 Spear Phishing Attack | 5 |
| 1.2.3 URL Attack..... | 5 |
| 1.2.4 Search Engine Phishing Attack..... | 5 |
| 1.2.5 Drive-by-download Attack..... | 6 |
| 1.3 Problem Statement..... | 6 |
| 1.4 Research Questions | 7 |
| 1.5 Goal and Objectives | 7 |
| 1.6 Motivation..... | 8 |
| 1.7 Contribution and Significance of Research | 8 |
| 1.8 Limitations of The Study | 9 |
| 1.9 Thesis Outline..... | 9 |
| Chapter Two: Background and Literature Review | 10 |
| 2.1 Overview | 11 |
| 2.2 Ensemble Classification Techniques | 11 |
| 2.3 Literature Review | 13 |
| 2.4 Summary..... | 29 |
| Chapter Three; Methodology and the Proposed Model..... | 35 |
| 3.1 Overview | 36 |
| 3.2 Methodology | 36 |

| | |
|---|-----------|
| 3.3 Collecting Dataset | 39 |
| Chapter Four: Implementation and Evaluation Results | 43 |
| 4.1 Introduction..... | 44 |
| 4.2 Experimental Parameters..... | 44 |
| 4.2.1 Random Forest | 45 |
| 4.2.2 Support Vector Machine | 45 |
| 4.2.3 Decision Tree | 46 |
| 4.2.4 Proposed Method | 47 |
| 4.3 Performance Evaluation..... | 48 |
| 4.3.1 Correctly and Incorrectly Classified Instances | 51 |
| 4.3.2 Kappa Statistic | 53 |
| 4.3.3 Mean Absolute Error | 54 |
| 4.3.4 Root Mean Squared Error | 55 |
| 4.3.5 Relative Absolute Error..... | 56 |
| 4.3.6 Root Relative Squared Error | 57 |
| 4.4 Confusion Matrix Comparison Between Models | 57 |
| Chapter Five: Conclusion and Future Work..... | 60 |
| 5.1 Conclusion..... | 61 |
| 5.2 Future Work..... | 61 |
| References | 63 |

List of Tables

| Chapter Number. Table Number | Contents | Page |
|---|---|-------------|
| 2.1 | Literature review summary | 30 |
| 3.1 | Features and description of Input Site List | 41 |
| 4.1 | Experiment Parameters for Random Forest | 45 |
| 4.2 | Experiment Parameters for SVM | 46 |
| 4.3 | Experiment Parameters for Decision Tree | 46 |
| 4.4 | Experiment Parameters for The Proposed Module | 47 |
| 4.5 | Comparative Analysis Between Existing and proposed Model | 49 |
| 4.6 | Weighted average of Confusion Metric Comparison Among Learning Models | 57 |

List of Figures

| Chapter Number. Figure Number | Contents | Page |
|----------------------------------|--|------|
| 2.1 | General Ensemble Architecture | 12 |
| 3.1 | Proposed Methodology | 38 |
| 3.2 | WEKA GUI interface | 40 |
| 4.1 | Comparative Analysis of Results | 50 |
| 4.2 | Correctly Classified Instances Graph | 51 |
| 4.3 | Incorrectly Classified Instances Graph | 52 |
| 4.4 | Kappa Statistic Graph | 53 |
| 4.5 | Mean Absolute Error Graph | 54 |
| 4.6 | Root Mean Squared Error Graph | 55 |
| 4.7 | Relative Absolute Error Graph | 56 |
| 4.8 | Root Relative Squared Error Graph | 57 |
| 4.9 | Weighted Average of Confusion Metric Comparison Among Learning Models | 58 |

List of Abbreviations

| Abbreviation | Meaning |
|--------------|---|
| AC | Associative Classification |
| APT | Advanced Persistent Threat |
| APWG | Anti-Phishing Working Group |
| ARM | Association Rule Mining |
| ARFF | Attribute-Relation File Format |
| AUC | Area Under Curve |
| DT | Decision Tree |
| EDRI | Enhanced Dynamic Rule Induction |
| FACA | Fast Associative Classification Algorithm |
| FN | False Negatives |
| FP | False Positives |
| FST | Feature Selection Technique |
| HTTPS | Hyper Text Transport Protocol security |
| IG | Information Gain |
| KNN | k-Nearest Neighbours |
| LST | Least Square Twin |
| MAE | Mean Absolute Error |
| MCAC | Multi-label Classifier based Associative Classification |
| MCAR | Multiple Classification based on Associative Rules |
| ML | Machine Learning |
| MLP | Multi-Layer Perceptron |
| MSE | Mean Square Error |
| NN | Neural Network |
| OBIE | Ontology-Based Information Extraction |
| PII | Personally Identifiable Information |
| RAE | Relative Absolute Error |
| RMSD | Root-Mean Square Deviation |
| RRSE | Root Relative Squared Error |
| SU | Symmetrical Uncertainty |
| SVM | Secure Virtual Machine |
| TN | True Negatives |
| TP | True Positives |
| UCI | University California Irvine |
| URL | Uniform Resource Locator |

How To Detect Phishing Website Using Three Ensemble Classification

Prepared By

Yussra M. AL-Shareef

Supervisor

Dr. Hesham Abusaimh

Abstract

As the number of web users increases, phishing attacks are gradually increasing. In order to effectively respond to various phishing attacks, a proper understanding of phishing attacks is necessary, and appropriate response methods must be utilized. To this end, in this thesis, three ensemble classification to detect the phishing website attack is analyzed. Through this analysis, it is possible to reconsider the awareness of phishing attacks and prevent the damage of phishing attacks in advance. In addition, a countermeasure is proposed for each phishing type based on the analyzed content. The proposed countermeasure is a method that utilizes appropriate website features for each step. To determine the effectiveness of the countermeasure, every classification model is generated through the proposed feature extraction method and the accuracy of each model is verified. In conclusion, the proposed method in this thesis is the basis for strengthening anti-phishing technology and the basis for strengthening website security. Therefore, ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance bagging or improve prediction stacking. Phishing website detection algorithm using three ensemble classification, which is proposed in this thesis can get the high phishing website detecting accuracy, because three classification algorithms Random Forest, Support Vector Machine, and Decision Tree are combined in one system. All the achieved proposed algorithm results have shown the highest accuracy of 98.52% than others. It is higher 1.26% than Random Forest, 3.16% than Support Vector Machine, and 2.65% than the Decision Tree algorithm.

Keywords: Phishing website, Support Vector Machine, Decision Tree, Random Forest, machine learning, Three Ensemble, Classification.

كيفية اكتشاف موقع التصيد الاحتيالي باستخدام تصنيف المجموعات الثلاثية

إعداد

يسرى ماجد الشريف

إشراف

الدكتور هشام ابو صايمة

المُلخص

مع زيادة عدد مستخدمي الويب ، تزداد هجمات التصيد الاحتيالي تدريجياً. من أجل الاستجابة بفعالية لمختلف هجمات التصيد الاحتيالي ، من الضروري الفهم الصحيح لهجمات التصيد الاحتيالي ، ويجب استخدام طرق الاستجابة المناسبة. تحقيقاً لهذه الغاية ، في هذه الاطروحة ، يتم تحليل تصنيف المجموعة ثلاثية النماذج للكشف عن هجوم موقع التصيد. من خلال هذا التحليل ، من الممكن إعادة النظر في الوعي بهجمات التصيد ومنع حدوث هجمات التصيد مقدماً. بالإضافة إلى ذلك ، يُقترح إجراء مضاد لكل نوع تصيد بناءً على المحتوى الذي تم تحليله. النموذج المقترح هو طريقة تستخدم ميزات موقع الويب المناسبة لكل خطوة. لتحديد فعالية الإجراء المضاد ، يتم إنشاء كل نموذج تصنيف من خلال طريقة استخراج الميزات المقترحة ويتم التحقق من دقة كل نموذج. في الختام ، فإن الطريقة المقترحة في هذه الاطروحة هي الأساس لتقوية تكنولوجيا مكافحة التصيد والأساس لتعزيز أمن الموقع. لذلك ، تعد طرق تصنيف المجموعة الثلاثية النماذج تجمع بين العديد من تقنيات التعلم الآلي في نموذج تنبئي واحد من أجل تقليل التباين في التباين أو تحسين التنبؤ. يمكن لخوارزمية اكتشاف مواقع التصيد باستخدام ثلاث تصنيفات للمجموعات ، والتي تم اقتراحها في هذه الأطروحة أن تحصل على دقة عالية في اكتشاف مواقع التصيد ، لأن ثلاث خوارزميات تصنيف هي والغابة العشوائية و دعم آلة المتجه و شجرة القرارات مدمجة في نظام واحد. أظهرت جميع نتائج الخوارزمية المقترحة التي تم تحقيقها أعلى دقة بنسبة 98.52% عن غيرها. وهي أعلى بنسبة 1.26% خوارزمية الغابة العشوائية. ، و 3.16% من دعم آلة المتجه ، و 2.65% من خوارزمية شجرة القرار

الكلمات الرئيسية: موقع التصيد الاحتيالي ، دعم آلة المتجه ، شجرة القرار ، الغابة العشوائية ، التعلم الآلي ، المجموعة الثلاثة ، التصنيف.

Chapter One

Introduction

1.1 Introduction

The majority of financial and public institutions have recently upgraded and enhanced the direct online services provided to their customers. In that regard, America and other developed countries in Europe still continuously using online shopping. As the number of Internet-based services increases, technology has led to the spread of smartphones, their increasing use has seen huge groups of people who depend more and more on online services such as shopping, online banking, settling their bills, or even playing games with friends and strangers. These activities have led to an effect on the universal economy, and a great dependency on online financial services which has increased the security risk for clients as well as financial institutions(Fortune Magazine, 2011).

Crime also occurs online, such as phishing, which is crime centered around identity theft. There are many stories and incidents in the media regarding groups that target customers by phishing. In order to protect customers, financial institutions have tried to improve online safety, as fraudsters are constantly evolving their style of attack. Phishing websites are maliciously created to mimic real-world webpages (Fortune Magazine, 2011). The phisher usually creates webpages which visually resemble real webpages with the intention of defrauding the victim. For example, a customer who is unaware of this type of fraud can be easily deceived. In this scenario, the phishing victim's webpage on their device will display their bank account, passwords, credit card numbers, or other confidential information to the owners of phishing webpages. Although phishing is a comparatively newer crime in comparison to other online crimes such as viruses and piracy, there have been noticeable increases in the amount and intensity of phishing incidents across the world (Aburrous, Hossain, Dahal, and Thabtah, 2010).

The objective of a phishing website is gaining personal information without permission, either by blackmail or through visiting an imitation webpage that resembles the real one, which requests that the user enters personal information. This results in information security breaches through compromises in confidential data whereby the victim might suffer a financial or asset loss. The attacker may additionally commit identity theft using the personal details of victims. Also, a phishing attack can harm the reputation of the financial institution which has been spoofed, as customers lose confidence that their account is secure. Consequently, they may take their custom to another company. Phishing, if not investigated, can negatively impact an organization's assets, revenue, customer relationship, or marketing effort, as well as their corporate image. A phishing attack might cost company hundreds of thousands of dollars for each attack in terms of personnel time and fraud-related loss. Additionally, costs linked to harm to consumer confidence and brand image can reach millions of dollars (Brooks, 2006).

Regarding definition, the term phishing originates from digital crimes relying upon email bait to phish for passwords and other personal or confidential information. The concept is that bait is thrown out in the hope that users will bite, just as a fish does. The bait can be an e-mail or instant message, which via a link takes the users to a phishing website (James, 2006). Because of the many types of data which are captured, both management efficiency and rapid retrieval of information are vital when making decisions. Data mining is the extraction of information from a vast dataset. Data mining or knowledge discovery methods are used in various areas, including financial analysis, decision support, industrial retail, and market analysis (Ayesha et al., 2010).

1.2 Types of phishing Attacks

Phishing is a type of security attack where the phishing is a criminal technology that uses both social and technological techniques to steal a malicious site or steal information by installing a malicious program on a user's PC to steal the user's personal information or financial account proof information tempts the victim through a fake website to voluntarily reveal personal details (Ming and Yang, 2006). The fisher here impersonates or act as a: banker, online tradesman, on credit card company. (Seker, 2006).

Therefore, an appropriate phishing response plan is required. To study effective countermeasures against phishing, it is necessary to have a clear understanding of the phishing process and to analyze several phishing websites attacking detection algorithms. As the success rate of phishing scams increases, phishing is gradually becoming intelligent in various types. Here are some of the most common ways in which they target people

1.2.1 Clone Phishing Attack

Clone phishing attack is an attack that attracts people by creating a homepage similar to a legitimate homepage that actually exists. A type of attack that involves phishing by replicating websites that users visit frequently. these sites usually ask users for login information. The replicate website stores the user's information on the attacker's server for use in future attacks. In some cases, it is classified as a web spoofing attack. Modern web browsers have built-in security indicators that protect users from phishing, such as domain names and HTTPS. However, many cases of damage occur because they are ignored by careless users (Nazreen Banu, Munawara Banu 2013).

1.2.2 Spear Phishing Attack

Spear phishing is an attack that targets employees of a specific institution or company and induces access through e-mail or other methods. It is a type of Advanced Persistent Threat (APT) attack. In order to induce a user's click, it is often disguised as a similar organization sending mail. When an email attachment is executed, an attack that leads a user to a malicious code distribution site is executed, or a malicious code is directly executed to infect the user's PC. According to TREND MICRO, 91% of targeted attacks start with spear phishing emails, and 94% of spear phishing emails are attached files. Since 76% of the targets are companies or government agencies, the amount of damage is large (Nakashima, Harris, 2018).

1.2.3 URL Attack

This is an attack that can lead to a malicious site when a user clicks on a link disguised as a normal site. Attacks involving similar domain names or attacks using technically disguised links may be involved (Ubing et al. , 2019).

1.2.4 Search Engine Phishing Attack

It is an attack that leads a user to a phishing site by manipulating it to be ranked high when a user searches through a search engine vulnerability. An attacker creates a phishing site and allows search engines to rank phishing websites at the top. If an attacker masquerades as a normal site and provides a product that is of interest to customers, it can be registered in a search engine. Therefore, the search engine displays both the normal site and the phishing site when displaying the search results of the user. Users trust the search results of search engines, so they connect to phishing sites without a doubt. When a site is visited, a malicious program is installed, or the personal

information is provided to the phishing site through the membership registration process or through an attack disguised as product purchase information (Huh, Kim, 2012).

1.2.5 Drive-by-download Attack

When using the Internet through a web browser, it is an attack in which malicious code is automatically downloaded and executed without user consent by simply accessing the website. This attack exploits a vulnerability in a website, targeting popular software such as web browsers, Flash, and Java. A malicious script is executed due to a vulnerability in the software, and the malicious code is downloaded and executed due to the script (Irwin,2020).

1.3 Problem Statement

Phishing websites are fake websites that can be constructed by attached to imitate and represent legitimate websites to cheat other people through stealing their personal vital data such as bank accounts information, national insurance number, passwords. (Ubing et al. ,2019). Therefore, the results will breach of information security via the theft of confidential data where the victim incurs a financial loss. In brief, it is online fraud or delinquency to the highest degree (Abur-rous,Ragheb,2011). Consequently, the assessment or discovering of phishing websites requires an intelligent model enabling the recognition and detection of the suspicious features related to phishing websites.

The main problem addressed in this study is the strengthening of user authentication on the internet website. The research investigates the potential uses of three ensemble classification models in detecting phishing websites. In particular, the aim here is the development of an ensemble model that will be used for predicting whether a website is phishy or legitimate, and if so to what degree, to improve the detection accuracy of phishing websites.

1.4 Research Questions

To attempt addressing the limitations discussed in the previous section, this thesis is aimed to answer the following research questions:

- Is the classification of data mining, particularly three ensemble helps more to predict phishing?
- Which rule based classification technique is more accurate in predicting phishing websites?
- Is the classification of data mining, particularly ensemble ones, useful tools to predict phishing?
- What are the other information sources used or required for identifying fraudulent websites?
- When a phishing website is identified, how can the user be informed?

1.5 Goal and Objectives

The goal is the building of an intelligent phishing detection model that uses data mining methods, an ensemble, to assess if phishing activity is occurring on a website. The resultant implementation must be effective and practical, can provide accurate identification, for instance, the avoidance of false negatives and positives, and be able to inform the user clearly about the phishing risk rate of the website being visited. We are also developing a real-time browser add-on that will provide warnings when visiting suspicious sites.

The research has the following objectives:

- The launch of an extensive and critical study focusing on the aspects of phishing and ensemble data mining techniques.
- The development of new ensemble data mining algorithms for the website

phishing issue.

- Conducting a comprehensive empirical study to evaluate the proposed models on different phishing collections such as the Anti-Phishing Working Group repository.

1.6 Motivation

A phishing website, as a process, is a complicated issue to analyse and understand because it includes social and technical problems for which there is no silver bullet to directly solve it. This is why all phishing website factors and characteristics are processed quantitatively and qualitatively to understand where to concentrate protective measures for the prevention or mitigation of every threat and risk stemming from a visit to phishing websites, particularly creating the trust crises which severely affects all online transactions.

The motivation or the aim of this study is having a resilient and effective model of intelligence for detecting phishing websites in assessing if phishing activities are occurring, to help every user from the catastrophic consequences of having their passwords and personal information stolen.

1.7 Contribution and Significance of Research

We propose an effective detection system that crawls websites and automatically discovers malicious pages. We intend our system to be used by a blacklist provider who can automatically compile and maintain an up-to-date blacklist of malicious Uniform Resource Locator (URLs). Our system is equipped with a plentiful set of features that reflect various types of essential characteristics of the webpage content or behavior, which are impossible or difficult to be camouflaged by the miscreants. We focus on characterizing the nature of such websites using only the information from the website

and training a machine learning classifier to distinguish between phishing and legitimate websites. Consequently, the contribution of this research work is the employment of the abovementioned three phases which differs from other research work such as (Nagaraj, Bhattacharjee, and Sridhar, 2018; Ubing, Jasmi, Abdullah, Jhanjhi, and Supramaniam, 2019) that only employed two phases to predict phishing websites.

1.8 Limitations of The Study

The proposed descriptor is limited to deals with texts and cannot treat or deal with other forms. The study is comparative and limited to the use of dataset for Using three ensemble classification to detect phishing websites.

1.9 Thesis Outline

Introducing detect phishing websites using three ensemble classification gives an overview of the proposed model and types of phishing attacks. The research problem, research questions, goal and objectives, motivation, contribution, and significance of the research, and limitations of the study are also discussed. The rest of this thesis is organized as follows:

Chapter Two discusses the literature review on detect phishing website and its shortcoming.

Chapter Three discusses in detail a description of proposed algorithm.

Chapter Four presents the implementation of the proposed descriptor. The results and its effectiveness are also discussed in this chapter.

Chapter Five will give a general summary of the thesis, summarizes the research findings and future works.

Chapter Two
Background and Literature
Review

2.1 Overview

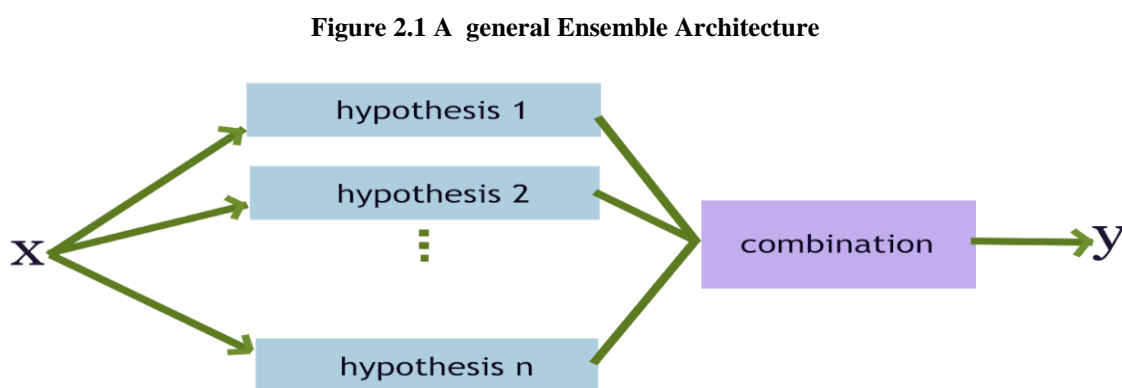
This chapter presents literature reviews on proposals on many anti-phishing techniques are presented to reduce phishing attacks through prevention and detection. The concept of ensemble learning is an ensemble of algorithms that use more than one learning model. Section 2.2 discusses the related work and presents overall comparison between related work. Section 2.3 provides a summary of this chapter.

2.2 Ensemble Classification Techniques

One of the major tasks of machine learning algorithms is to construct a fair model from a dataset. The process of generating models from data is called learning or training and the learned model can be called as hypothesis or learner. The learning algorithms which construct a set of classifiers and then classify new data points by making a choice of their predictions are known as Ensemble methods.

It has been discovered that ensembles are often much more accurate than the individual classifiers which make them up. The ensemble methods, also known as committee-based learning or learning multiple classifier systems train multiple hypotheses to solve the same problem. One of the most common examples of ensemble modeling is the random forest trees where a number of decision trees are used to predict outcomes.

Figure 2.1 shows a general Ensemble Architecture(Zhou, and Zhi-Hua ,2012):



An ensemble contains a number of hypothesis or learners which are usually generated from training data with the help of a base learning algorithm. Most ensemble methods use a single base learning algorithm to produce homogenous base learners or homogenous ensembles and there are also some other methods that use multiple learning algorithms and thus produce heterogeneous ensembles. Ensemble methods are well known for their ability to boost weak learners.

Some of the commonly used ensemble techniques three major kinds of meta-algorithms are discussed below (Zhou, and Zhi-Hua ,2012):

- **Bagging**

Bagging or Bootstrap Aggregation is a powerful, effective and simple ensemble method. The method uses multiple versions of a training set by using the bootstrap, i.e. sampling with replacement and it can be used with any type of model for classification or regression. Bagging is only effective when using unstable (i.e. a small change in the training set can cause a significant change in the model) non-linear models.

- **Boosting**

Boosting is a meta-algorithm which can be viewed as a model averaging method. It is the most widely used ensemble method and one of the most powerful learning ideas. This method was originally designed for classification, but it can also be profitably extended to regression. The original boosting algorithm combined three weak learners to generate a strong learner.

- **Stacking**

Stacking is concerned with combining multiple classifiers generated by using different learning algorithms on a single dataset which consists of pairs of feature vectors and

their classifications. This technique consists of basically two phases, in the first phase, a set of base-level classifiers is generated and in the second phase, a meta-level classifier is learned which combines the outputs of the base-level classifiers.

2.3 Literature Review

Recently, there are many investigations and researches related to the phishing detection model that uses data mining methods, where the use of Using three ensemble classification. In this chapter, some of the developed studies will be discussed.

Nagaraj, Bhattacharjee, Sridhar, and Sharvani (2018) stated that there was a lack of available techniques for detecting phishing activity and avoiding deception. They stated that the classification of phishing and non-phishing web content is an important issue in any security information protocol. However, fool-proof methods have not been implemented in practice. Therefore, the aim of the study is the presentation of an ensemble machine learning model for phishing website classification. Experimental simulations were conducted, and the performance of the ensemble model was compared with other machine learning algorithms. Additionally, a set of comparisons was conducted among several machine learning classifiers. In their study it was found that the random forest algorithm initially achieved better prediction accuracy of 93.41% compared to all the other machine learning algorithms which were tested. Furthermore, since the random forest algorithm performed best in detecting phishing websites, it was included in the twofold ensemble model, together with feedforward neural network, bagging and boosting neural networks, to produce a predictive model that is accurate and reliable in the classification of unknown data instances.

Ubing et al., (2019) focused on participating in developing the accuracy of phishing website detection. Accordingly, a feature elicitation algorithm was selected and

combined with an ensemble learning approach, which depends on plurality voting and parallels with a variety of classification models inclusive of Random Forest, Logistic Regression, Prediction model, etc. The study determined that present phishing detection methods have an accuracy rate between 70% and 92.52%. The experimental simulation verified that the accuracy rate of our suggested model can return to 95%, which was greater than the present methods for phishing website detection. Furthermore, the learning models have been used through the experiment determined that their suggested model has a rising accuracy rate and can be recognized as the result in the experiment execute through Azure, especially trees. To label the overfitting problem while adjust to expanding the indicator accuracy, the suggested solution model used feature extraction and ensemble learning where multiple learning models were decomposing to outcome a prediction. They used multiple models, the prediction was not viewpoint towards one model and was in place of depending on a greater number of predictions such that all predictions from any model effect the final ensemble prediction.

Abdel Hamid, Ayesha, and Thabtahb (2014) developed an Associative Classification (AC) method termed the Multi-label Classifier based Associative Classification (MCAC) to examine if this method is applicable for the detection of website phishing, and subsequently to test its accuracy. To achieve this, they identified the differentiating features of phishing websites from legitimate ones; they also surveyed intelligent ways for handling the phishing issue. In their research, they proved higher accuracy and better ability of AC, particularly MCAC, in detecting phishing websites in comparison to other intelligent algorithms. Additionally, AC data mining methods were used to identify feature interrelationships and present them in a simple yet effective control. The developed method enables the discovery of new rules that are combined with at least two classes. This gives users new types of instructions which are useful in comparison

to other intelligent approaches. They enhanced the criteria of the classification accuracy in determining the phishing websites based on their obtained experimental results. Their intention is using the test websites as training data after they have been classified, making the phishing model incremental.

In another study Discuss AC, Hadi, Aburub, and Alhawari, (2016) presented a new AC algorithm called the Fast-Associative Classification Algorithm (FACA). In their study, this algorithm was tested against four well-known AC algorithms including CBA, CMAR, Multiple Classification based on Associative Rules (MCAR), and ECAR. Their comparison was mainly based on classification accuracy and F1 evaluation measures. The results obtained from this research indicated that the FACA excelled and outperformed the other four algorithms in both the F1 and the accuracy evaluation measures. Moreover, another result from this research highlighted the fact that there is a potential for the prediction of phishing websites by means of using computerized data mining techniques.

Some researches were conducted not only on a single method of detecting phishing but took on multiple models and compared their performances.

Abdelhamid, Thabtah, and Abdel-Jaber, (2017) explored in their article the Machine Learning (ML) techniques that available to detect phishing attacks and define their advantages and disadvantages. Especially, different variants of ML techniques have been investigated to inform the fitting options that can operate as anti-phishing tools. Basically, they experimentally analyzed large numbers of ML techniques on real phishing datasets and pertaining results to different metrics. The target of their comparison was to explain the advantages and disadvantages of ML predictive models and to display their real performance when it comes to phishing attacks. They found out that the experimental simulation that displayed cover path models are more applicable

as anti-phishing solutions, specifically for beginner users, because of their simple yet effective knowledge bases in addition to their good phishing detection estimates. Lately, the most active way to combat phishing that depends on machine learning techniques has appeared. In this method, certain patterns were extracted by an ML technique and were used to classify websites either as legitimate or phishing, depending on certain features. The aim of this study was to define which ML approach is most effective in detecting phishing attacks by using a real dataset of 11,000 phishing websites. To achieve this aim, large numbers of ML methods have been compared with estimates to different metrics, inclusive features into the status and its effect on the phishing detection rate. Bayes Net and Support Vector Machine (SVM) have showed good performance with an estimate of accuracy. However, their models were hard to understand by end-users. On the contrary, Enhanced Dynamic Rule Induction (EDRI) and Ridor algorithms seemed to be appropriate for achieving high accuracies and being easy to understand. In the near future, the aim to combine an SVM within a web browser and order live experiments using huge numbers of users in a pilot study.

Mohammad ,Thabtah , and McCluskey, (2014) tried to find a solution for the phishing problem by means of using a self-structuring neural network, due to the neural networks need to have their structures constantly improved in order to cope with the constantly changing features that are significant in determining the type of web pages. Thus, automation of the process of structuring the network has solved this problem effectively. This model displayed high approval for noisy info, fault tolerance, and high indicator accuracy. Many experiments were handled in their research, and many periods differ in each observation. From their experiment, they found that all produced structures have high judgment intelligence. It is well known that a good anti-phishing tool should estimate the phishing attacks in a good time scheme. They considered that

the opportunity of a good anti-phishing tool at a good time scheme is also important to increase the scale of predicting phishing websites and have found that this tool should be improved regularly through continuous retraining. Furthermore, they have found that the process of finding the best structure was very difficult, and in most cases, this structure was defined by trial and error. Therefore, an anti-Phishing model was figured out, and in case, for any reason, it needs to be updated, then this design will ease this process. Although the design architecture used in their research was kind of difficult, its rule was the usage of an adaptive scheme with four structures: structural simplicity, learning rate adaptation, structural design adaptation and early stopping technique based on validation errors. Although many algorithms planned to robotize the neural network design, most of them use a selfish scheme in determining the original structure by adding a new layer to the network or adding a new neuron to the hidden layer. The main idea behind this design was to spotlight on an adaptive scheme for both learning rate and network structure. The adaptive scheme is more comfortable because it can handle different positions that might exist during the designing phase. One of the future developments of this design was by adding a procedure to determine the significance of the features before they are approving in building a neural network-based anti-phishing system. In addition, they were outlining to create a toolbar that implements the design and combines it with a web browser. This toolbar should be up to date continuously to get by with any development on the weights, and in case, a new design was being reconstructed.

A study was conducted by Qabajeha, Thabtah and Chiclana (2018) that handled the comparison of the conventional methods with the technological methods of combating phishing websites. The conventional methods indicate the enforcing of cyber laws, and prosecuting phishers and malicious website creators. That is in addition to raising

awareness for the end-users about phishing websites and giving them certain indicators on how to detect them. On the other hand, the problem of phishing websites can be mitigated by implementing technological solutions to detect phishing websites, by means of using machine learning algorithms to detect and classify phishing websites. Such algorithms can be implemented in web browsers and warnings about phishing websites can be communicated to the end-user. Mainly, the algorithms discussed in this research were rule-based algorithms, decision trees, SVM, Neural Network (NN), and computational intelligence. It compared their performances, advantages, and disadvantages.

Bahnsen et al. (2017) suggested a method that was more effective for detecting phishing websites in real-time. Stated that there are a lot of anti-phishing methods appearing, but phishers use various and dynamic methods to fraud victims, so a smart and flexible model was needed to catch the phishing websites. Data mining methods could be used to promote an active model with the nontrivial and underlying data that could be a reserve from huge datasets using classification algorithms to label websites legal. Four different classification algorithms were utilized to classify the data set and approximately studied for their achievement, accuracy, and several criteria. The experiments were handled using four different rule-based algorithms to detect the hidden awareness, from the huge dataset to expect the phishing websites. Classified outcomes were parallel for their performances in the scheme of accuracy, error rate, time duration and the total number of criteria composed. However, the results showed that all the chosen algorithms complete higher expected rate. The rules were developed showed the interaction and relationship between website features and that can help us in creating phishing website detection frameworks. There was a phishing detection model

that is good to keep users from being phished by achieving verification through a private information submission.

Preethi, and Velmayil (2016) proposed another method to analyse the phishing URL's using lexical analysis. suggested the Pre-Phish algorithm which is a computerized machine learning to resolve phishing and non-phishing URL to outgrowth safe result. The phishing URLs mostly have a twosome of connections between the part of the enrolled domain level and the way or reservation level URL. Therefore, applying these connections URL is describing by inter-relatedness and it classifies using features extract from attributes. Also, these features after that used in the machine learning method to catch phishing URLs from an actual dataset. The classification of phishing and non-phishing website has been achieved by discovering the range value and threshold value for each attribute using decision-making classification. This technique was also classified in Mat lab using three main classifiers SVM, Random Forest, and Naive Bayes to detect how it is doing on the dataset estimate. This paper suggested the Pre-Phish algorithm to get an active phishing URL detection system depends on URL phrasal analysis. The approach of the Pre-Phish was an experimental phishing, an experimental case study that has been achieved to gather and evaluate the range of variety of phishing website features and patterns, with all its related attributes. This was a computerized machine learning technique that depends on attributes of phishing URL properties to catch and block phishing websites and to provide high-level security. The limitation of the work the same technique was used to establish a tool depending on a web browser add-on component which can catch and block phishing websites on actual time and resolve data mining approaches to detect new patterns of phishing URL.

Going further with rule-based algorithms, Thabtah, and Kamalov (2017) seriously tested the recent research studies on the use of expected models with constraint for

phishing detection and decide the capability of these methods on phishing. To achieve their task, they experimentally checked four different criteria-based classifiers that belong to selfish, associative classification and criteria induction methods on real phishing datasets and with respect to multi evaluation measures. However, they evaluated the classifiers copied and comparing them with known classic classification algorithms including Bayes Net, and Simple logistics. The purpose of the contrast was to indicate the advantages and disadvantages of the expected portrait with criteria and declare their real performance when it comes to detecting phishing activities. The results surely viewed that EDRI is the newest selfish algorithm that not only achieves useful portrait but also is high performing with respect to expected accuracy as well as runtime when they are selected as anti-phishing tools. They had one approach to reduce the danger associated with phishing was to create automated expected models using rule-based classification techniques. To accomplish this purpose, rule-based classifiers that apply to a multi-group of algorithms have been used (RIPPER, EDRI, RIDOR) along with other two non-rule classic classification algorithms (Probabilistic-Bayes Net, Simple logistic). The bases of relation were indicating error rate, time-consuming to create the expected models in minutes and what does the model contains. In addition, they have also taken characteristic filtering into the examination and its response to the phishing detection rate. The experimental simulation against huge phishing websites informed that the rule-based classifier was a highly useful anti-phishing technique, after all, they derived balanced size models without holding up the expected accuracy performance. In the real-world, the criteria detected by EDRI and RIPPER algorithms, are strong in differentiating websites, since they can distribute as decision tools for end-user to attack phishing. Moreover, the limitation they have was the aim to create rule cut

back approaches to further decrease the number of rules derived by rule-based expected forms.

In this article, Aburrous, Hossain, Dahal, and Thabtah (2010) discussed a novel technique to take the deadlock and complexity in identity and predicting the e-banking phishing website. They suggested an intelligent flexible and active model that depends on using cooperative and classification Data Mining algorithms. These algorithms were used to describe and detect all the element and criteria in order to categorize the phishing website and the relation that connect them with each other. they achieved six variety of classification algorithms and approaches to determine the phishing training data sets rules to categorize their legitimacy. Also, they correlated their performances, accuracy, a total of criteria achieves and speed. A Phishing Case study was tested to create the website phishing process. The criteria developed from the associative classification model viewed the correlation between some critical features such as URL and domain Identity, and security and encryption rule at the end of the phishing detection rate. The experimental simulation establishes the utility of using AC approaches in actual operation and its better performance in comparison to other common classifications algorithms. Moreover, for future study, they aimed to use variety of shortening methods such as lazy pruning which will cancel criteria that falsely categorize training items and manage all other criteria to be used by MCAR associative classification technique orderly to reduce the size of the appearing classifiers and to temporarily degree and analyses the effect of these various clipping on the final analysis.

As a form of another approach that suggested by Al-diabat (2016) who discussed the exploration of features elicitation proposes to detect the active set of features in the scheme of classification performance. he made a comparison of two known features

elicitation approach orderly to detect the minimum set of features of phishing selection using data mining. Experimental result on a massive number of features data set has been completed using Information achievement and connected Features set approaches. additionally, two data mining algorithms labelled as C4.5 and IREP have been tested on various sets of detected features to display the advantage and disadvantage of the feature detection operation. In addition, he had the ability to detect new observations in the forms of criteria that display critical connection during important features. Therefore, detecting the most important features for the website's phishing trouble was the main task for both security and data mining experts. Also, in this paper, the author measured two popular feature detection methods namely Symmetrical Uncertainty (SU) and Information Gain (IG) assuming various features and defining small sets of connections through features. This is important for reducing the uncertainty correlated with phishing and may help in creating new anti-phishing results. Moreover, the outlines have two common data mining techniques to measure the importance of features on two rules: phishing detection rate and classifier size. In another concept, tested selfish and decision tree algorithms on various versions of an actual security dataset correlated to phishing. Finally, in the future, the author will develop the opportunity to combine the target of known feature detection approaches to enhance the accuracy of the solution of the pre-processing stage.

Nandhini, and Vasanthi (2017) discussed how features are extracted to help classify phishing websites using the above-mentioned algorithms. The authors reviewed the features of detection the purpose is to detect the valid set of features in the schema of categorizing performance. In order to compare the features detection and categorize technique orderly to detect the bottom set of features of phishing selection using data mining. Experimental result was a massive number of features data set has been

completed using data growth and connection Features set approaches. Moreover, five data mining algorithms; Naïve Bayes, k-Nearest Neighbors (KNN), Random Forest, SVM and j48 have been used to categorize the web phishing data set, analyses the results and detect the performance approach to categorize the web page phishing data set.

Information categorizing is a critical application area in web mining and web page phishing data sets why because categorizing billions of phishing transcripts annually it is costly and a time-wasting task. Then, the automated classifier is created using pre-classified fragment phishing data set whose accuracy and time efficiency it is rather than annual classification and expectation. Detecting an efficient model also shows the main criteria in text classification. Data mining classification methods request to be created to be active controlling huge numbers of items with different numbers. Essentially, all the known methods for classification like decision trees rules, Bayes methods and SVM classifiers have been used to the state of phishing data. In this research study, a web page's phishing data sets were used to develop the different classification methods and find out the active classifier. They made a comparison between this information by presenting the material to the conventional method of Bayesian statistical classification, J48 Decision tree, Random Forest, KNN and SVM to form a classification pattern. The Random forest model shows better performance than KNN, SVM, J48, and Naïve Bayes classification patterns. Future works may also contain hybrid classification models by linking some of the web mining approaches like attribute detection and clustering.

Varshney, Misra, and Atrey (2016) worked on the avoidance, detection, and education of phishing aggression, but to date, they stated that there was no complete and accurate result for preventing them. This research tests and classified the most important and

novel approaches suggested in the area of phished website detection and outlines their advantages and drawbacks. Additionally, an accurate investigation of the newest theory suggested by authors in kinds of subcategories was produced. In addition, this article indicated the advantages, drawbacks, and research differences in the area of phishing website detection that could be treated upon in future research and evolution. The result given in this article will help academia and production to find the best anti-phishing approach. In this article, it was suggested the techniques for phishing detection have been taking. In this study attracted on the case that phishing detection plane executed rather than phishing avoidance and user training result because they do not inform modification in verification stage and do not depend on the user's ability to detect phishing. Additionally, the phishing detection result is cheaper than phishing avoidance results in a schema of the more hardware required and password administration. This article classified phishing detection results in six classifications and displays the advantages and drawbacks of using any one of them. Also, they identified that search engine-placed techniques are the clear available result for phishing detection as they only require a single search engine reservation result with its critical algorithm to detect phishing websites at the user's end. And It can be extending both at the client-side or on the server-side, SEB techniques need neither machine learning nor training. There platform absolute and can be extended over any browser and over any operating system as a browser add-on. And they had many threats in the area of search engine dependent phishing detection, like developing phishing detection accuracy when a long-term decent range determine to start carrying out nasty phishing activity; and decreasing the number of false positives for decent domains that are working for a very short period of time and are therefore not viewed with the top search results.

On the contrary, Rathod, Kapse (2017) stated that different anti-phishing approaches make use of various features of the webpage in order to detect the fraud website. In this study, they explained a variety of phishing detection approaches and displayed the survey of different phishing website detection techniques. In this study, they have evaluated and displayed different phishing detection approaches. Various approaches use different properties of the web page like URL, text, security certificates, host information, etc. In order to catch phishing websites, and the approaches displayed in this work have variant defects in the scheme of accuracy and performance. There was no individual system that can catch all kinds of phishing attacks therefore in the future, there needs to be an all-in-one entity that will catch all these attacks with high accuracy and performance. Most of the approaches still have a constraint in the scheme of zero-hour attack, fixed objects in web pages, computational power, accuracy, and performance. All these points require to be solved in the future.

On the other side of this spectrum, Patil, and Devale (2016) discussed the different techniques of phishing attacks. The authors presented a violation of testing phishing by developing the technical devices and social engineering to effort the incur option of unfamiliar users. This technique often spread an accurate organization so as to control a user to perform a plane if request by the mimicked entity. Most of the time, phishing aggression is being recorded by the exercised users, but security is the main motivation for beginner users as they are not familiar with such resources. However, some techniques are limited to look after phishing aggression only and the problem in selection is essential. Proposed to underline the different methods used for the selection of phishing aggression. They have also detected different methods for the selection and prevention of phishing. Nevertheless, isolated from that, they have presented a new design for the selection and prevention of phishing aggression. Phishing could not be

determined with a single result. It is an important position in which Phishers usually try to occur with label new approaches to managing the user. Online users should lock formal risk reviews to determine the newest method which may head to developing Phishing aggression. To get a secure path, the user must be familiar are about the risks of leading to malware which is catching place nowadays. Further improvement is complete in selection the identity steals and the phishing emails. It does not include the growing aim of e-mail deploy. In other words, they can also have said that other electronic performance will also get a part of the challenge. And they are proposed to really work on this trouble before aggression is being caught wildly. A request should be informed which can defect all critical internet banking operations.

Mahalakshmi, Goud, and Murthy (2018) discussed the types of phishing and the resulting conflicts of it. The authors stated that phishing is one of the general engineering approaches that collect special information via websites like wicked websites and ambiguous e-mail to request personal information from a corporation or an individual by jump as a convincing entity or organization. Phishing often aggression email by using as a coach and even transfer messages by email to users that display a few of a company or an institution that execute business like a financial institution, banking, etc. Furthermore, they stated that the phishing is becoming more fraudulent day by day and its selection is very critical. In cyberspace, phishing is prompt the scientist to analyses the model during which they can improve more security towards the secure services produced by the web. This paper also explained the kinds of phishing and competition need to it. This report will advise the general public for catching avoidance as well as careful steps across the phishing attack. As the internet is one of the most present phishing aggression by message so the anti-phishing lacks to be concerned for these which have been used by a number of people. It is a review of the

phishing aggression inform to be responded by anti-phishing by supplying the data about the phishing forward with it against measures for anti-phishing methods.

Mande and Thosar (2018) discussed in their article the Phishing website which looks forward to picking the victim's private information by distracting them to wave a fake website page that like a real to quality one is another kind of offender law through the internet and its one of the especially involvement toward various areas including e-managing an account and huckster. The Phishing website detection was really an unexpected and piece issue including various items and rules that are not stable. On account of the previous and in addition to the vagueness in organizing sites because of the intelligent techniques programmers are useful, some intensely exciting strategies can be helpful and powerful tools can be applied, such as fuzzy, neural system and data mining approaches, which can be a successful structure in characteristic phishing websites. They have defined properties of phishing aggression and thus, they suggested a model in order to the classification of the phishing aggression. Their model consists of feature expression from websites and classification section. In the feature extraction, and they determined criteria of phishing feature extraction and these criteria have been used for access features. Moreover, they should also have trained for every user not to widely follow the links to websites where they must enter their personal information and that it is crucial to check the URL before getting on the website.

Shetty, and Niranjana (2016) defined in their article the concept of transgression in cyber security because of phishing message was detected from present messages which were sent over so net site social networking sites, these transgressions motivate to an explosion in network connection and steal of Personally Identifiable Information (PII) that causes a number of point-like identity crime and cyber fraud. To explain these problems, a system used advanced Ontology-Based Information Extraction approach

(OBIE) and Association Rule Mining (ARM) named as Anti Phishing Detection System that detect and then expect the phishing activity by managing frequently restore phishing database which contains of information gathered from previous attack to crush security; and block the phishing activity to support the user data. Gross will send uneasy messages through cell phones, and So net sites, which is difficult to continue their criminal activity powerfully. After surveying various structural patterns of mobile phones, present messengers and so net sites', it assisted to establish a new platform, which fighting phishing by using the rule-mining and Ontology methods to perfectly classify and to suppose phishing violation. When messages are detected phishy, then the details of criminal are outlined, and the victim is informing with certain kinds of challenge activity. As a future work, Phishing messages can be detected on a governmental level to create a robust so net sites and detection should be complete if ambiguous messages are transferred using multimedia format.

Specifically, Shrivas and Suryawanshi (2017) discussed the usage of decision tree classification of phishing websites. The authors defined the guarantee of the data is a very threatening task for every organization and institute to enhance the order of data and connected technology. According to them, phishing aggression is one of the most critical points across private data from an illegal person. Data mining depends on classification intelligent approaches to do very critical criteria to categorize phishing and non-phishing aggression. In this study work, they suggested decision tree technique and Info Get Feature Selection Technique (FST) using various top detecting feature subdivisions for analyzing computationally active models for classification of phishing websites. Furthermore, they suggested the Decision Tree method allows the best classification accuracy as 99.80% with 15 numbers of features in the state of Info gets FST. A phishing attack is a very serious problem for internet users and face by e-mail

users. Classification is a critical approach is used to detect and categorize of phishing and non-phishing aggression.

2.4 Summary

Table 2.1: Literature Review Summary.

| No. | Authors | Problem | Solution | Result |
|-----|---|---|--|--|
| 1. | Nagaraj, Bhattacharjee, Sridhar, and Sharvani, (2018) | Intrusion detection that nullifies phishing attacks | Classify phishing websites using ensemble twofold model using attributes for classification. | Random Forest produced a high accuracy compared to previously used algorithms of 93.41per cent |
| 2. | Ubing et al., (2019) | evaluating whether a website is legitimate or phishing | analyze phishing and non-phishing URLs to produce real result | with all its relations produced a high accuracy compared to previously used algorithms of 92.52% |
| 3. | Abdel Hamid, Ayesha, and Thabtah, (2014) | Investigating phishing websites using the AC model | Developing AC into MCAC to provide improved and more accurate results | The MCAC algorithm was shown to outperform RIPPER, C4.5, PART, CBA, and MCAR with 1.86%, 1.24%, 4.46%, 2.56%, 0.8% 1.24%, 4.46%, 2.56%, 0.8% |
| 4. | Hadi, Aburub, and Alhawari, (2016) | Well known AC algorithms have low accuracy in detecting phishing websites | Developing a new AC model called FACA (Fast associative classification algorithm) | The classification accuracy was reduced for FACA, CBA, CMAR, MCAR And ECAR by only 0.04%, 0.02% 0.04%, 0.07%, 0.06% |
| 5. | Abdelhamid, Thabtah, and Abdel-Jaber, (2017) | Comparing different ML algorithms and finding advantages | Comparing large numbers of ML techniques on real phishing datasets | Covering approach model are more appropriate as anti-phishing solutions. The accuracy produced |

| No. | Authors | Problem | Solution | Result |
|-----|---|---|---|---|
| | | and disadvantages of each | | between 90% to 96% |
| 6. | Mohammad, Thabtah , McCluskey, (2014) | Phishing website techniques evolve rapidly, and detection algorithms need to evolve accordingly to cope with them and stay up to date | Developing a self-structuring neural network that adapts to the changes of phishing techniques using automatic machine learning | All produced structures have high generalization ability and have high accuracy of 94.07% |
| 7. | Qabajeha, Thabtahband Chiclanaa,(2018) | Conventional approached to combat phishing such as raising awareness are not as effective | Proposing a technological-based method to combat phishing using machine learning algorithms | AC methods generated more rules than the rest of the algorithms and the accuracy 83% |
| 8. | Bahnsen et al. ,(2017) | The phishing attacks its increased | Investigate the use of URLs as the input in machine learning model applied for phishing site prediction | Evaluate the performance of the feature based on URLs lexical and statistical analysis then trained a random forest classifier and the accuracy rate of 93.5% |
| 9. | Preethi, and Velmayil, (2016) | Phishing is fraudulent Technique achieved by phishing web page | Introduce the pre phish algorithm which is an automated machine learning approach to improving the accuracy of phishing website models including Random forest, Logistic Regression, Prediction model detection. Employed the algorithm was selected and integrated with an ensemble learning | That implemented to gather and analyze range of different phishing website features Prove the accuracy rate which is higher than the current technology for phishing website detection And the accuracy rate 97.83% |

| No. | Authors | Problem | Solution | Result |
|-----|--|--|--|---|
| | | | methodology, and compared with different classification | |
| 10. | Thabtah, and Kamalov, (2017) | Recent research studies using predictive models that are not as effective at phishing detection | Evaluating four different rule-based classifiers | EDRI generates useful models which are highly competitive with respect to predictive accuracy, C4.5-Rules achieved 0.86%, 3.03%, and 3.33% higher percentages of accuracy than RIPPER, RIDOR and EDRI algorithms respectively |
| 11. | Aburrous, Hossain,Dahal,and Thabtah,(2010) | use semantic attacks for targeting used instead of computers. It is quite a new internet crime compared to other forms | Used a unique approach for overcoming the difficulties and complexities in the detection and prediction of e-banking phishing websites | Demonstrated the appropriateness of Associative Classification techniques in a real application and their improved performance in comparison to other traditional classifications algorithms, with an accuracy of 88.4 % |
| 12. | Al-diabat,(2016) | Phishing is a problem that mimicking legitimate websites to deceive online users in order to steal their sensitive information | investigates features selection aiming to determine the effective set of features in terms of classification performance | two data mining algorithms namely C4.5 and IREP have been trained on different sets of selected features to show the advantage and disadvantage of the feature selection and the accuracy 96.5 |
| 13 | Nandhini ,and Vasanthi,(2017) | The problem is the phishing website to steal sensitive information | investigates features selection aiming to determine the effective set of features in terms of classification | tests on large number of features data set have been done using IG and correlation features set methods, the |

| No. | Authors | Problem | Solution | Result |
|-----|--------------------------------------|--|--|---|
| | | | performance | accuracy 92.98% |
| 14. | Varshney, Misra, and Atrey,(2016) | The problem is the phishing website to steal sensitive information | investigates features selection aiming to determine the effective set of features in terms of classification performance | identify the best anti-phishing technique and the accuracy is 97.16%. |
| 15. | Rathod,Kapse,(2017) | The problem is the phishing website to steal sensitive information | These websites look exactly like the original website | discuss different phishing detection techniques and present the survey of various phishing website detection approaches. and provides accuracy in terms of true positive and false positive |
| 16. | Patil, and Devale, (2016) | The problem is the phishing website to steal sensitive information | investigate many techniques used for detection of phishing attacks. And discovered various techniques for detection and prevention of phishing | introduced a new model for detection and prevention of phishing attacks. |
| 17. | Mahalakshmi, Goud, and Murthy,(2018) | phishing attackers in the means to abuse the personal details of clients | Develop more security towards the safe service provided by the web | Discuss types of phishing and conflicts due to it and have highest accuracy. |

| No. | Authors | Problem | Solution | Result |
|-----|-------------------------------------|---|---|---|
| 18. | Mande and Thosar, (2018) | This is a Web phishing attack, which is the major problems in web security | Developing an algorithm of (ELM) extreme learning machine | Used IP address and URL Age of Domain, Non-coordinating URLs to present how easy to use the classifier as a feature of the evaluation function with classification accuracy respectively |
| 19. | Shetty, and Niranjana, (2016) | The violation in the cyber security lead to disturbance in network communication and larceny of personal identifiable information (PII) that causes plenty of issues like identity theft and cyber scam | a system is developed using Ontology based Information Extraction technique (OBIE) and Association rule mining (ARM) named as Anti Phishing Detection System | specifies the computation time taken to identify phishing words using Data mining and WordNet Ontology. Proposed system identifies the phishing words faster than keyword-based approach, the accuracy is 75%. |
| 20. | Shrivastava and Suryawanshi, (2017) | Phishing attack is one of the important issues to access the sensitive information from unauthorized person | proposed decision tree technique and IG feature selection technique (FST) using different top selected feature subsets for developing computationally efficient model for classification of phishing websites | Decision Tree (DT) technique gives better classification accuracy as 99.80% with 15 numbers of features in case of IG FST and the accuracy for each classifier Decision Tree (DT)91.80%, Random Tree 66.75%, Random Forest 78.85%, Decision Stump 84.73 |

In this chapter, most of the research works that have been presented used single data mining classifier with training and validation to detect phishing website. In addition,

there are some of them used the multiple classifier to detect the phishing website. Moreover, they have been a bit slow and non-accurate in determining the phishing website. Therefore, these methodologies need to have better approach using multi layered classifier becomes required to detect the phishing website fast and accurately. In fact, using a single classifier in the field of machine learning may lack robustness the performance on the training and validation when applied in real-life situations such as phishing detection problem. Hence, it is necessary to have a new intelligent data mining algorithm that combines multiple classifiers in order to increase the performance prediction. Combining multiple classifiers plays an important role in enhancing the accuracy of the classification process, in addition to allowing decision makers to easily identify the legitimate and illegal website.

Chapter Three

Methodology and the Proposed Model

3.1 Overview

This chapter presents the proposed model of building an intelligent phishing detection model, which uses data mining methods, an ensemble, to assess if phishing activity is occurring on a website.

3.2 Methodology

This research work attempts to evaluate different machine learning techniques that aim to investigate the potential uses of three ensemble classification models in detecting phishing websites. In particular, the aim here is the development of an ensemble model that will be used for predicting whether a website is phishy or legitimate, and if so to what degree. At this stage determining the phishing website can be viewed as a data mining classification problem, wherein this instance the class attribute is the degree of phishing. The classification process is based upon attributes and characteristics which are used to distinguish phishy sites such as spelling mistakes, long URLs, personalization, prefixes, and suffixes. These attributes are obtained from input websites using various tools. It must be considered here that the firm belief of the author is that the results of this thesis will open a new door for research paths in the area of predicting and detecting phishing websites using ensemble or data mining methods, where there are many potential domain applications, particularly e-banking which can invest in and profit from it. Therefore, an intelligent three-step ensemble learning model to predict phishing websites will be designed and developed. Figure 3.1 depicts the general framework of the proposed phishing prediction methodology.

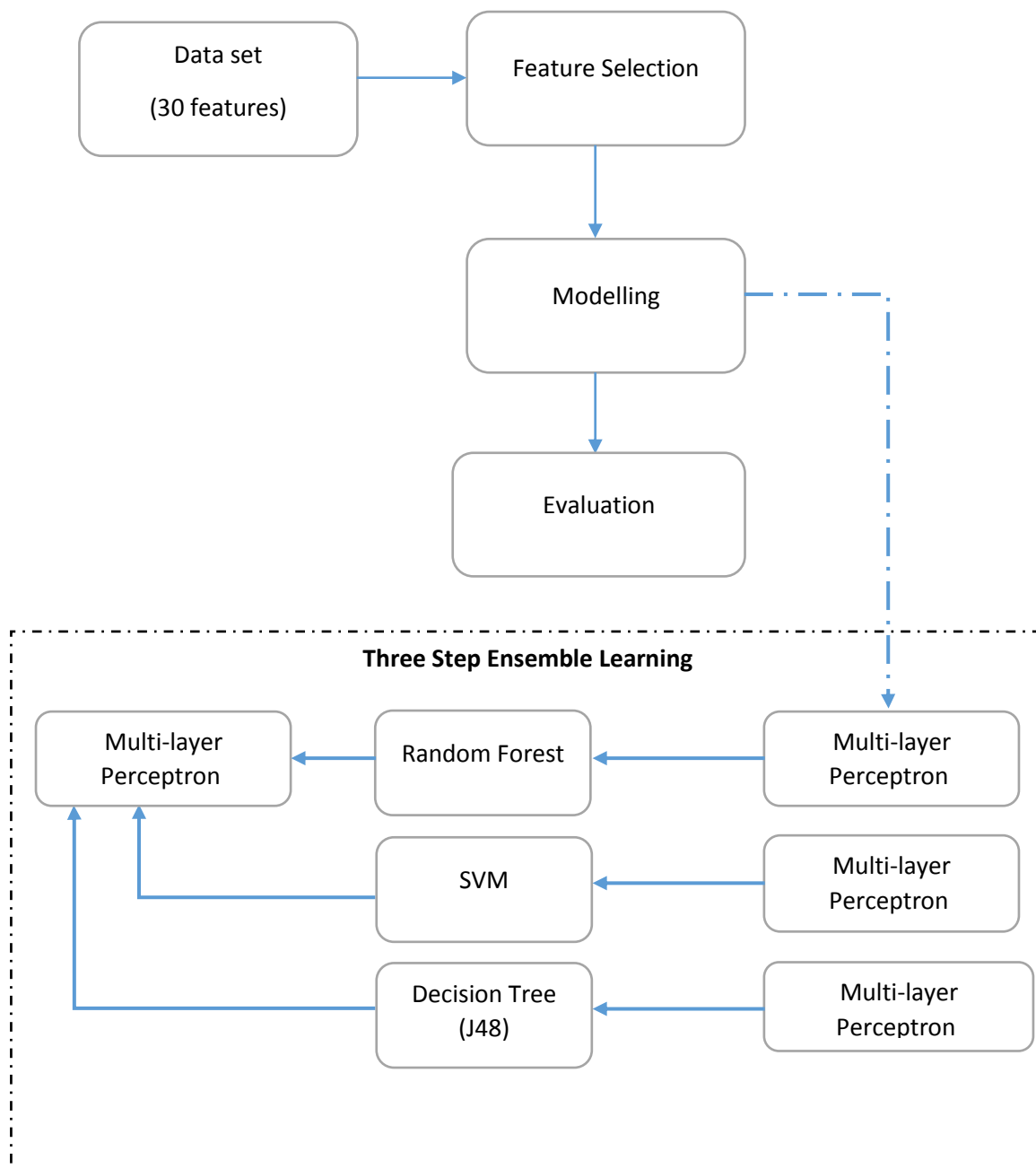


Figure 3.1: Proposed Methodology.

This methodology consists of three main phases, namely feature selection, modelling, and evaluation. In the feature selection phase, the chi-square feature selection method will be employed on the inputted data set from University California Irvine (UCI) dataset which can be used in the comparisons that will be conducted between this research work and the already conducted comparisons in this set. The module is implemented to extract the features from the input site. In the proposed model illustrate the

association rule mining algorithms on a phishing URL data set, in ARFF format, from UCI machine learning repository. the data set is relatively balanced containing 11055 instances, 4898 phishing, and 6157 legitimate, each instance has 30 features. This phase aims to select to the most significant features such as text, URL, log data, and more to distinguish between legitimate and phishing websites. While in the modelling phase, a three-step ensemble framework will be developed to handle the selected most significant features from the first phase. Weka which is a collection of machine learning algorithms for data mining will be used to develop such framework. The first step aims at combining three different multi-layer perceptron neural networks to work concurrently. While in the second step, the Random Forests methodology are a will be applied which is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest , Decision tree forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest , and SVM Is machine learning algorithm that analyses data for classification and regression analysis. SVMs are used in text categorization, image classification, handwriting recognition and in the sciences will be applied on the resulted information from the previous first step. Finally, a single multi-layer Perceptron neural networks will be applied on the resulted information from the second step.

The final phase is the evaluation phase which aims to assess the overall performance of the suggested classification framework, therefore, the most widely applied evaluation metrics for phishing detection problems such as classification accuracy, sensitivity, specificity, g-mean, F1 evaluation, precision, recall and Area Under Curve (AUC) will be applied in this phase. Consequently, the contribution of this research work is the employment of the abovementioned three phases which differs differ from other

research work such as (Nagaraj, Bhattacharjee, & Sridhar, 2018; Ubing, Jasmi, Abdullah, Jhanjhi, and Supramaniam, 2019) that only employed two phases to predict phishing websites.

I will implement to extract the features from the input site. In the proposed model illustrate the association rule mining algorithms on a phishing URL data set, in ARFF format, from the University of California Irvine UCI machine learning repository. the data set is relatively balanced containing 11055 instances, 4898 phishing, and 6157 legitimate, each instance has 30 features. And use Microsoft Excel to view the results.

3.3 Collecting Dataset

This section describes the properties and lists some statistics about the utilized datasets.

Weka 3.8.4 is a Java based open source software created by the University of Waikato

University of New Zealand and has the following GUI interface as shown Figure 3.2.



Figure 3.2: WEKA GUI interface.

The ARFF is an ASCII text file that describes a list of instances sharing a set of attributes. ARFFs were developed by the Machine Learning Project of the Faculty of

Computer Science at the University of Waikato for use with Weka machine learning software. In this thesis ARFF format dataset is used.

The module is implemented to extract the features from the input site. In the proposed model illustrate the association rule mining algorithms on a phishing URL data set, in ARFF format, from the University of California Irvine UCI machine learning repository. the data set is relatively balanced containing 11055 instances, 4898 phishing, and 6157 legitimate, each instance has 30 features as show in Table 3.1.

Table 3.1: Features and description of Input Site List (Aburrous, et al., 2008; Mohammad,Thabtah, and McCluskey. , 2014 ; Preethi ,and Velmayil, 2016).

| Srl. | Feature Name | Feature Description |
|-------------|----------------------------|--|
| 1 | having_IP_Address | Using an IP address in the domain name of the URL. |
| 2 | URL_Length | Length of URL Phishers can use long URL to hide the doubtful part in the address bar. |
| 3 | Shortining_Service | URL shortening service is a third-party website that converts that long URL to a short, case-sensitive alphanumeric code. |
| 4 | having_At_Symbol | The “@” symbol leads the browser to ignore everything prior it and redirects the user to the link typed after it. |
| 5 | double_slash_redirecting | The existence of “//” within the URL path means that the user will be redirected to another website. |
| 6 | Prefix_Suffix | Phishers try to scam users by reshaping the suspicious URL, so it looks legitimate. One technique used is adding a prefix or suffix to the legitimate URL. Thus, the user may not notice any difference. |
| 7 | having_Sub_Domain | Another technique used by phishers to scam users is by adding a subdomain to the URL so users may believe they are dealing with an authentic website. |
| 8 | SSL final_State | is a standard security technology for establishing an encrypted link between a server and a client. |
| 9 | Domain_registration_length | Based on the fact that a phishing website lives for a short period of time, we believe that trustworthy domains are regularly paid for several years in advance. |
| 10 | Favicon | graphic image icon associated with a |

| Srl. | Feature Name | Feature Description |
|-------------|---------------------|--|
| | | specific webpage |
| 11 | port | This feature is useful in validating if a particular service such as HTTP is up or down on a specific server. In the aim of controlling intrusions, it is much better to merely open ports that you need. |
| 12 | HTTPS_token | IF The phishers may add the HTTPS token to the domain part of a URL in order to trick users. |
| 13 | Request_URL | If the objects are loaded from a domain other than the one typed in the URL address bar, the webpage is potentially suspicious. |
| 14 | URL_of_Anchor | Similar to the URL feature, but here the links within the webpage may point to a domain different from the domain typed in the URL address bar. |
| 15 | Links_in_tags | Links present in tags like META and SCRIPT are checked |
| 16 | SFH | Server Form Handler contain an empty string or “about:blank” are considered doubtful because an action should be taken upon the submitted information. |
| 17 | Submitting_to_email | Web form allows a user to submit his personal information that is directed to a server for processing. A phisher might redirect the user’s information to his personal email. To that end, a server-side script language might be used such as mail() function in PHP. One more client-side function that might be used for this purpose is the “mailto: |
| 18 | Abnormal_URL | If the website identity does not match a record in the WHOIS database (WHOIS, 2011) the website is classified as phishy. |
| 19 | Redirect | Redirection is commonly used by phishers to hide the real link and lures the users to submit their information to a fake site. |
| 20 | on_mouseover | Phishers often hide the suspicious link by showing a fake link on the status bar of the browser or by hiding the status bar itself. This can be achieved by tracking the mouse cursor and once the user arrives to the suspicious link the status bar content is changed. |
| 21 | RightClick | Phishers use JavaScript to disable the right-click function, so that users cannot view and save the webpage source code. |
| 22 | popUpWidnow | Usually authenticated sites do not ask users to submit their credentials via a |

| Srl. | Feature Name | Feature Description |
|-------------|------------------------|--|
| | | popup window. |
| 23 | Iframe | is an HTML tag used to display an additional webpage into one that is currently shown |
| 24 | age_of_domain | Websites that have an online presence of less than 1 year, can be considered risky. |
| 25 | DNSRecord | An empty or missing DNS record of a website is classified as phishy. |
| 26 | web_traffic | Legitimate websites usually have high traffic since they are being visited regularly. Since phishing websites normally have a relatively short life; they have no web traffic or they have low ranking. |
| 27 | Page_Rank | PageRank is a value ranging from “0” to “1”. PageRank aims to measure how important a webpage is on the Internet. The greater the PageRank value the more important the webpage. In our datasets, we find that about 95% of phishing webpages have no PageRank. Moreover, we find that the remaining 5% of phishing webpages may reach a PageRank value up to “0.2”. |
| 28 | Google_Index | This feature examines whether a website is in Google’s index or not. |
| 29 | Links_pointing_to_page | In our datasets and due to its short life span, we find that 98% of phishing dataset items have no links pointing to them. On the other hand, legitimate websites have at least 2 external links pointing to them. |
| 30 | Statistical_report | formulate numerous statistical reports on phishing websites at every given period. |

Chapter Four

Implementation and Evaluation

Results

In this Chapter, implementation and evaluation results performed on detailed Attribute-Relation File Format (ARFF) dataset, which is used in the proposed algorithm is explained, and analyze results, comparing the proposed algorithm with Random Forest, Support Vector Machine (SVM), Decision Tree algorithms. Moreover, the research questions are answered in light of the research and its results.

4.1 Introduction

The proposed methodology uses the dataset with the maximum 30 inputs for the experimental study and generates the counter results. The proposed model also uses the three-step technique that provides the proposed system an edge over the existing methodologies to overcome the drawbacks.

After studying the literature surveys of the same, the proposed model concludes that below mentioned techniques are the best in collaboration, to detect the phishing attack performed on the websites:

- Random Forest.
- SVM.
- Decision Tree(J48).

4.2 Experimental Parameters

For legitimate comparisons, similar inputs were tested on each one of the three combined detectors which are Random Forest, SVM, and Decision Tree (J48) individually. These three detectors slightly varied in their results, however all of them scored less accuracy than the combined ensemble. Their individual results came as follows:

4.2.1 Random Forest

Random forest is an ensemble technique that is a combination of tree predictors. In which each tree is responsible for the unique output after feeding the independent random sample vector. Random forest is used for their error generalization technique, as the forest gets populated with number of trees; the accuracy of the random forest also increases. The accuracy totally depends on the correlation between the trees, after randomly selecting the features for error rate. The features used by the random forest could be generated by monitoring the error and correlation between nodes. This results in measuring the importance of variable. Following are the parameter used in the random forest shows in Table 4.1:

Table 4.1: Experiment Parameters for Random Forest.

| Srl. | Parameter | Value |
|-------------|---|--------------|
| 1. | Size of each bag | 100 |
| 2. | Number of Iteration | 100 |
| 3. | Number of execute slots | 1 |
| 4. | Number of attributes to randomly investigate | 0 |
| 5. | Minimum number of instances per leaf | 1 |
| 6. | Minimum numeric class variance proportion of train variance for split | 0.001 |
| 7. | Seed for random number generator | 1 |
| 8. | Number of cross-validation folds | 10-fold |

4.2.2 Support Vector Machine

SVM is a supervised machine learning technique for classifier builder. SVM aims to enable the prediction of labels by generating the decision boundary such as hyperplane in between the two selected classes from minimum one label. The hyperplane is responsible for the data points and the support vectors. It uses the distance of the data

points in such a way that each class can be classified separately. Following are the parameters used by the SVM for the experimental study shown in Table 4.2:

Table 4.2: Experiment Parameters for SVM.

| Srl. | Parameter | Value |
|-------------|---|----------------------------|
| 1. | The complexity constant | 1.0 |
| 2. | Use lower-order terms | 0.001 |
| 3. | The epsilon for round-off error | 1.0E-12 |
| 4. | The number of folds for the internal cross-validation | -1 |
| 5. | The random number seed | 1 |
| 6. | Number of kernel evaluations | 139775595 (69.611% cached) |
| 7. | Regulation parameter | 1.0 |
| 8. | Kernel Type | RBF |
| 9. | Gamma | auto |
| 10. | Number of support vectors | 1746 |
| 11. | Number of cross-validation folds | 10-fold |

4.2.3 Decision Tree

A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements. Following are the parameters used by the decision tree for the experimental study shows in Table 4.3:

Table 4.3: Experiment Parameters for Decision Tree.

| Srl. | Parameter | Value |
|-------------|--------------------------------------|--------------|
| 1. | Confidence threshold for pruning | 0.25 |
| 2. | Minimum number of instances per leaf | 2 |
| 3. | Number of Instances | 1105 |
| 4. | Number of Leaves | 169 |

| | | |
|----|----------------------------------|---------|
| 5. | Size of the tree | 297 |
| 6. | Number of cross-validation folds | 10-fold |

4.2.4 Proposed Method

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance bagging or improve predictions (stacking).

Phishing website detection algorithm using three ensemble classification, which is proposed in this thesis can get the high phishing website detecting accuracy, because three classification algorithms Random Forest, SVM and Decision Tree are combined in one system shows in Table 4.4.

Table 4.4: Experiment Parameters for The Proposed Module.

| Srl. | Parameter | Value |
|------|---|---------|
| 1. | Threshold ranking | -1.7977 |
| 2. | Number of attributes | 30 |
| 3. | Number of cross-validation folds | 10-fold |
| 4. | Learning rate for the backpropagation algorithm | 0.3 |
| 5. | Momentum rate for the backpropagation algorithm | 0.2 |
| 6. | Number of epochs to train through | 500 |
| 7. | Percentage size of validation set to use to terminate training | 0 |
| 8. | The number of consecutive increases of error allowed for validation testing | 20 |

| | | |
|-----|--|--------------|
| | before training terminates | |
| 9. | Gamma | auto |
| 10. | The value used to seed the random number generator | 1 -num-slots |
| 11. | Confidence threshold for pruning | 0.25 |
| 12. | Minimum number of instances per leaf | 2 |
| 13. | Size of each bag | 100 |
| 14. | Number of bag error | 100 |
| 15. | Number of execution slots | 1 |
| 16. | Number of attributes | 0 |
| 17. | Minimum number of instances | 1 |
| 18. | Minimum variance for split | 0.001 |
| 19. | Seed for random number generator | 1 |
| 20. | Sets the epsilon for round-off error. | 1.0E-12 |
| 21. | The exponent for the polynomial kernel | 1 |
| 22. | The complexity constant | 250007 |
| 23. | Set the maximum number of iterations | -1 |

4.3 Performance Evaluation

The first question which has to do with identifying the main characteristics of a phishing website can be answered according to the features it has which can be classified into four categories as worked on by (Mohammad, Thabtah, and McCluskey, 2014). The first category is Address-bar-based features. Which indicates as the name suggests that the address bar itself shows a suspicious or phishing website. Of what can be learnt about this category are those sub-types like using IP address in the address bar;

long URL to hide the suspicious part; shortening a URL; having an “@” sign in a URL; redirecting using the “//” sign; and more features that is shown in the address bar. The second category is the abnormal-based features. Abnormality is of many types such as request URL; URL of anchor; links in <meta>, <script>, and <link> tags; server form handler; submitting information to email; and abnormal URL. The third feature is based on HTML and JavaScript such as website forwarding; status bar customization; disabling right-click; using pop-up window; and iframe redirection. The last category of features is the domain-based, in which the phishing websites can be identified by age of domain; DNS records; website traffic; page rank; Google index; and other similar properties. The answer in numbers can be illustrated by the following Table 4.5:

Table 4.5: Comparative Analysis Between Existing and proposed Model.

| Detecting Method | Random Forest | SVM | Decision Tree | Proposed Model |
|---|----------------------|------------|----------------------|-----------------------|
| Correctly Classified Instances | 97.2592 % | 95.3596 % | 95.8752 % | 98.5256% |
| Difference compared to the Model | 1.2664% | 3.1660% | 2.6504% | 0% |

Namely, the three-step model is 1.2 % higher in accuracy than the Random Forest detector. It is 3.0996% more accurate than using the SVM alone. And finally, it is 2.584% higher in accuracy of detecting phishing websites than using Decision Tree individually. Thus, it is significantly effective in detecting phishing websites and more reliable than single detector model.

There is a number of ways to analyze the results of a predicted model but the most common factors that are included and considered by the researchers are mentioned below Figure 4.1 as shows the comparative analysis of existing algorithms and proposed model.

Commonly used evaluation measures including Recall, Precision, F-Measure, and Rand Accuracy are biased and should not be used without a clear understanding of the biases, and corresponding identification of chance or base case levels of the statistic. Using these measures, a system that performs worse in the objective sense of Unforcedness, can appear to perform better under any of these commonly used measures.

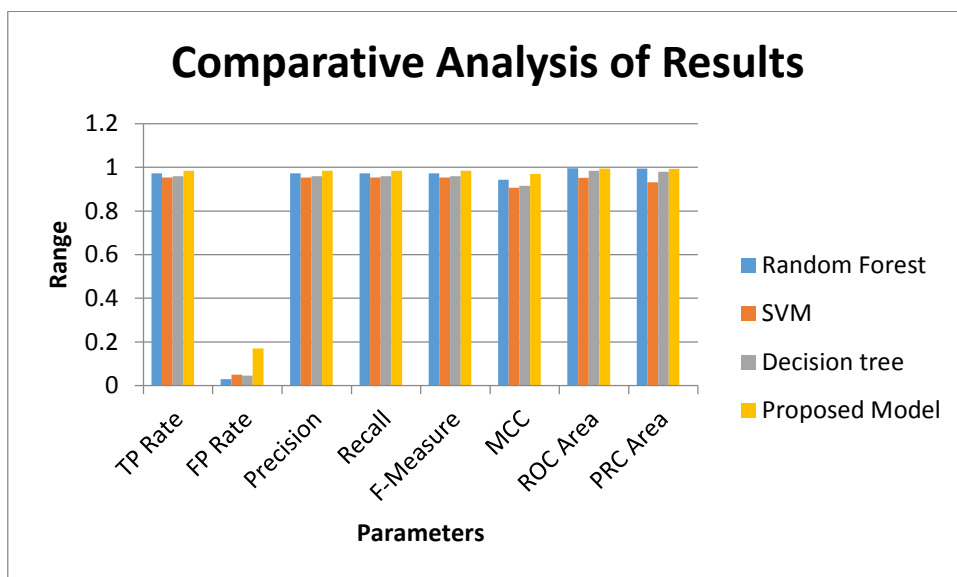


Figure 4.1: Comparative Analysis of Results.

As noted in Figure 4.1 shows the comparative analysis of Proposed Three-step model with other model as Random Forest, SVM and Decision Tree with different performance parameter. The value of precision has obtained better as compared to the original Random Forest, SVM and Decision Tree in the Basic mode. The best result for phishing website classification for Proposed with a precision of 98.52%, the second rank was for the original Random Forest with precision of 97.25%. The worst result was obtained by SVM and Decision Tree with precision of 95.35% and 95.87%. The precision enhancement of Proposed compared to the Random Forest is around 1.27% and the enhancement over the SVM and Decision Tree. The improved results of proposed can be justified due to the additional primitives such as lines that proposed

addresses. These results satisfy the objective of this thesis which aims at enhancing phishing website classification. The ROC curve and the confusion matrix of the Random Forests trained using their feature set are presented in Figure 4.2. It is evident that their ROC curve is worse than ours, with a smaller Area Under Curve (AUC). Meanwhile, the classifier trained using features presented in our thesis has higher TP and TN rate, and lower FP rate and FN rate.

4.3.1 Correctly and Incorrectly Classified Instances

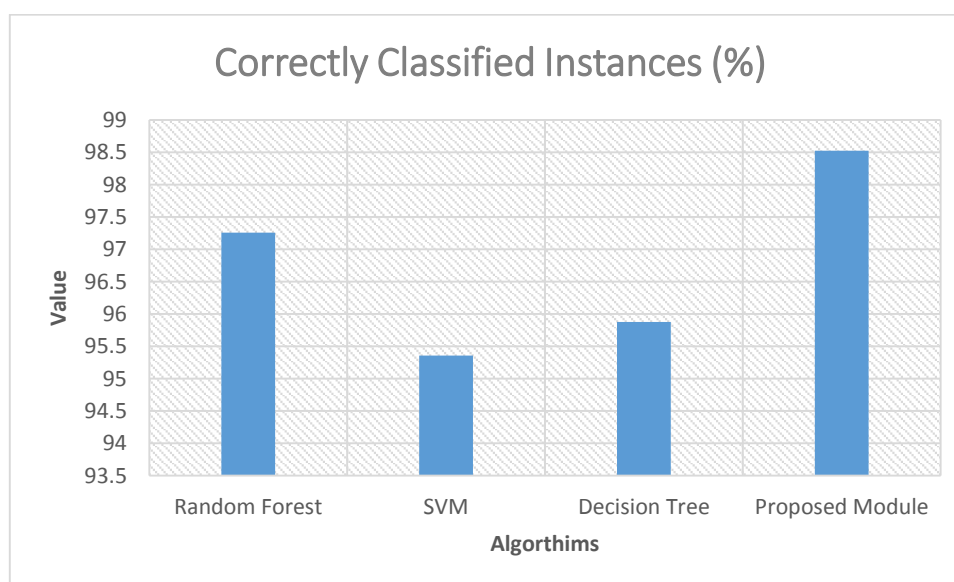


Figure 4.2: Correctly Classified Instances Graph.

In Figure 4.2, it shows that the performance of proposed model is much higher than that of other existing method on ARFF dataset. If we choose the best results among the results, the highest classification rate of the Proposed can reach 98.52% whereas other existing method gives lowest correctly classified rate 95.35%. The classification accuracy (%) for the contrasted algorithms derived from the phishing data.

Accuracy, which is refer to the ability of the algorithm to predict the correct class label for instances of unknown class labels (testing set), is calculated as given in Equation.

Accuracy measure is used for evaluating and comparing between the underlying descriptors (Ezziyyani, Bahaj, and Khoukhi, 2017).

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \dots\dots\dots 1$$

This Figure 4.3 is shown the percentage of correctly classified instances. The proposed algorithm has the highest accuracy 98.52% than others. It is higher 1.26% than Random Forest, 3.16% than SVM and 2.65% than Decision Tree algorithm.

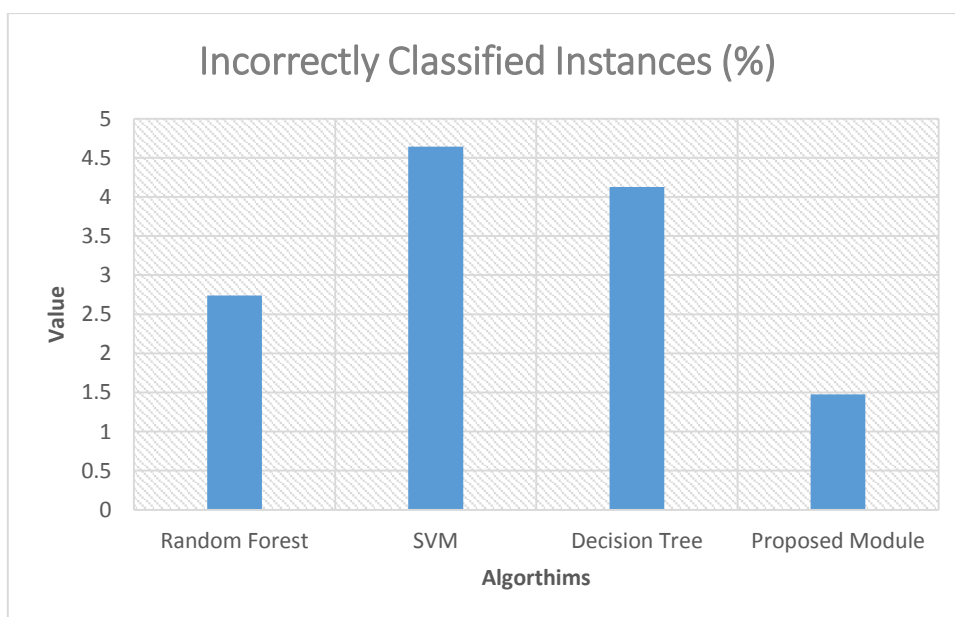


Figure 4.3: Incorrectly Classified Instances Graph.

Figure 4.3 is shown the percentage of incorrectly classified instances. The proposed algorithm has the lowest percentage 1.47% than others. It is lower 1.27% than Random Forest, 3.16% than SVM, and 2.65% than Decision Tree.

4.3.2 Kappa Statistic

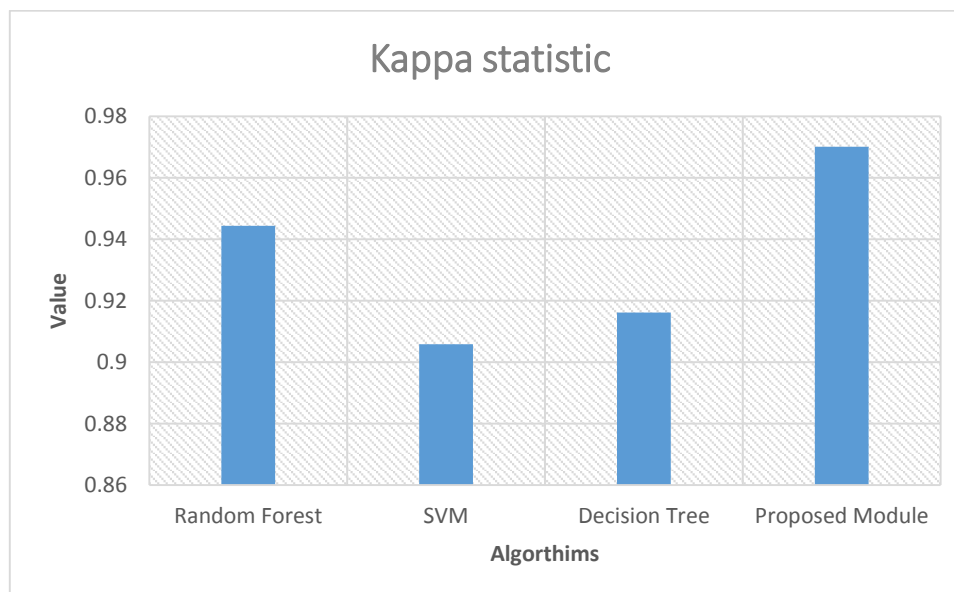


Figure 4.4: Kappa Statistic Graph.

Figure 4.4 is shown the kappa static, the proposed algorithm has the largest percentage of 97.01%, Random Forest 94.44%, Decision Tree 91.62%, and SVM has given the worst result of 90.58%. which is a statistic that is used to measure inter-rater reliability and also Intra-rater reliability for qualitative (categorical) items (McHugh, and Mary (2012)).

$$k = 1 - \frac{1 - P_o}{1 - P_e} \dots \dots \dots 2$$

Where equation number two P_o is the relative observed agreement among raters, and P_e is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If the raters are in complete agreement, then $k = 1$. If there is no agreement among the raters other than what would be expected by chance as given by P_e , $k \leq 0$. A kappa value of 0 means

that the result is the same as would be expected by chance (Gail, Benichou, Armitage, and Colton, 2000).

4.3.3 Mean Absolute Error

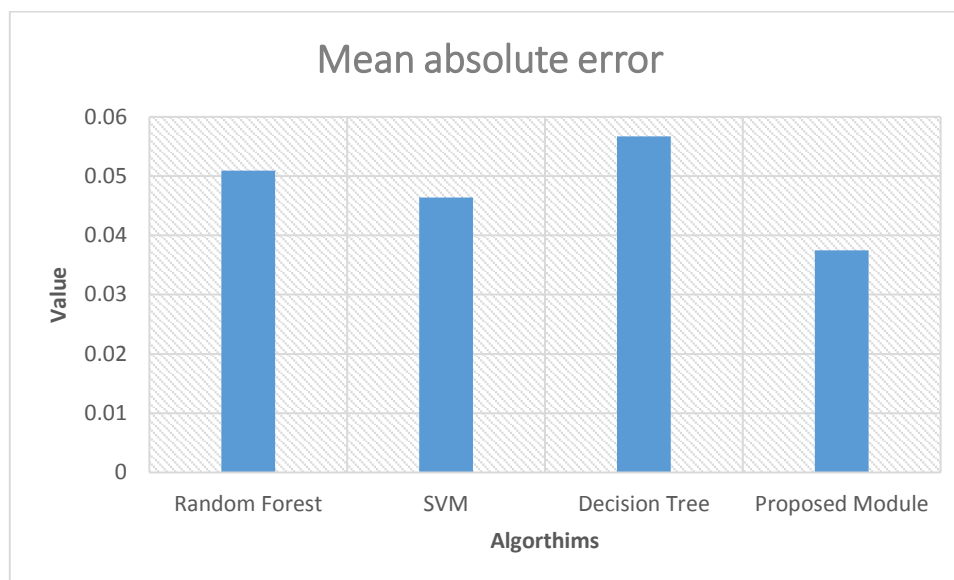


Figure 4.5: Mean Absolute Error Graph.

Figure 4.5 is shown the Mean Absolute Error (MAE), The proposed algorithm has the lowest percentage of 3.75% than others and Decision Tree has given worst result on MAE is 5.67%. which is a quantity used to measure how close forecasts or predictions are to the eventual outcomes.

$$\text{Mean Absolute Error} = \frac{1}{n} \sum_{i=1}^n |f_i - y_i| = \frac{1}{n} \sum_{i=1}^n |e_i| \text{ (Willmott, and}$$

Matsuura, 2005).....3

As the name suggests, the mean absolute error is an average of the absolute errors. $|e_i| = |f_i - y_i|$, where f_i is the prediction value and y_i is the true value. Note that alternative formulations may include relative frequencies as weight factors. The mean absolute error is like the variance, but rather than square the difference, use its absolute value. (If the scores are spread closely around the mean, the variance will be smaller than the mean absolute error. If the scores are not spread closely, squaring the distance

will lead to larger variances. Taking the absolute value assigns equal weight to the spread of data whereas squaring emphasizes the extremes. Squaring, however, makes the algebra easier to work with and relates to Pythagorean Theorem (Willmott, and Matsuura ,2005).

4.3.4 Root Mean Squared Error

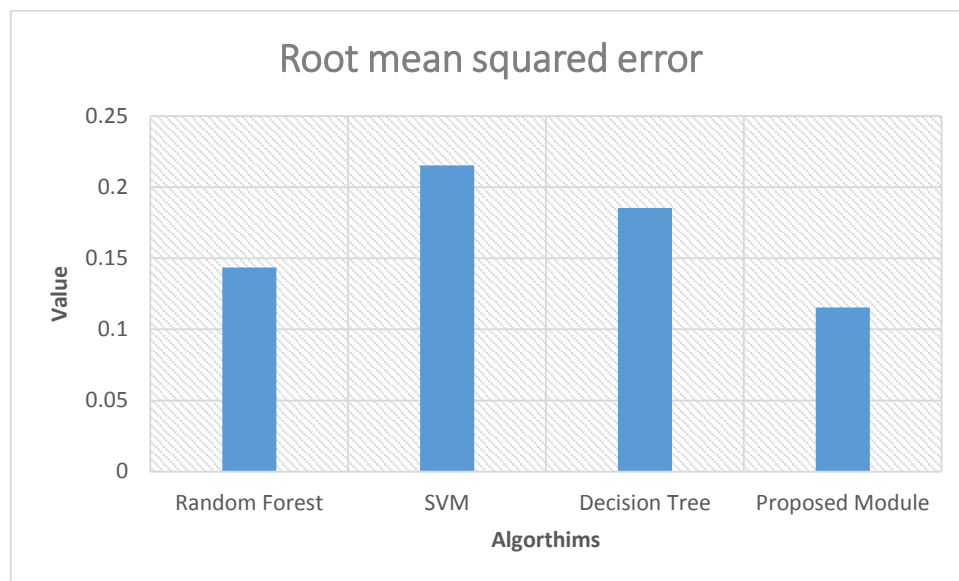


Figure 4.6: Root Mean Squared Error Graph.

Figure 4.6 is shown the percentage of the root mean squared error (RMSE). The proposed algorithm has the lowest percentage of 11.53% than others and SVM has given worst result on RMSE is 21.54%. Which is the measure of the differences between values sample and population values predicted by a model or an estimator and the values actually observed. It represents the sample standard deviation of the differences between predicted values and observed values. It aggregates the magnitudes of the errors in predictions for various times into a single measure of predictive power. It is a good measure of accuracy, but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale dependent. It is also called RMSE.

The RMSE of predicted values \hat{y}_t for times t of a regression's dependent variable y is computed for n different predictions as the square root of the mean of the squares of the deviations (Hyndman, and Koehler ,2006):

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y)^2}{n}} \dots\dots\dots 4$$

4.3.5 Relative Absolute Error

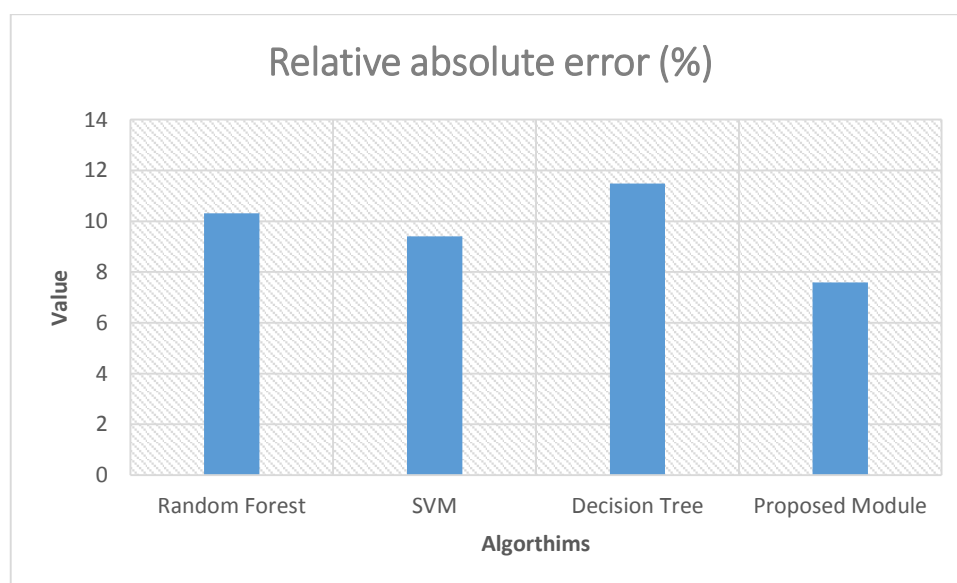


Figure 4.7: Relative Absolute Error Graph.

Figure 4.7 is shown the Relative Absolute Error (RAE). Absolute error is how much your result deviates from the real value. Relative error is a measure in percent compared to the real value. Figure 4.8 indicates the values of the RAE of 7.59% are lowest for proposed as compared to other methods and decision tree has given worst result on RAE is 11.48%.

4.3.6 Root Relative Squared Error

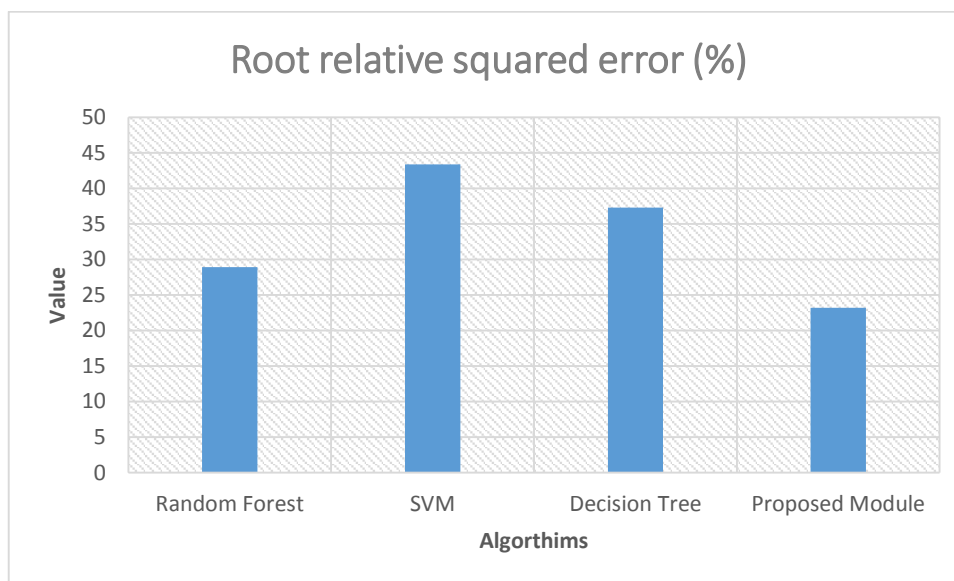


Figure 4.8: Root Relative Squared Error Graph.

Figure 4.8 is shown the Root Relative Squared Error (RRSE). The relative squared error normalizes the total squared error by dividing it by the total squared error of the simple predictor. The error is reduced to the same dimension as the quality being predicted by taking the square root of the relative squared error, RRSE of 0.11 was achieved from the test option of 10-fold cross validation. Figure 4.11 indicates the values of the RRSE of 23.2056 are lowest for proposed as compared to other methods and SVM has given worst result on RRSE is 43.3655%.

4.4 Confusion Matrix Comparison Between Models

Table 4.6: Weighted average of Confusion Metric Comparison Among Learning Models

| Srl. | Classification | True Positive | False Negative | False Positive | True Negative |
|------|----------------|---------------|----------------|----------------|---------------|
| 1. | Random Forest | 4705 | 193 | 110 | 6047 |
| 2. | SVM | 4591 | 307 | 206 | 5951 |
| 3. | Decision Tree | 4615 | 283 | 173 | 5984 |

| | | | | | |
|----|-----------------------|------|-----|-----|------|
| 4. | proposed Model | 4782 | 116 | 110 | 6047 |
|----|-----------------------|------|-----|-----|------|

This method required minimal user training and does not require any changes to the existing authentication schemes used by a website. The accuracy of the detection schemes is measured in terms of the following parameters:

Number of True Positives (TP): The number of phishing websites correctly labeled as phishing.

Number of True Negatives (TN): The number of legitimate websites correctly labeled as legitimate.

Number of False Positives (FP): The number of legitimate websites incorrectly labeled as phishing.

Number of False Negatives (FN): The number of phishing websites incorrectly labeled as legitimate.

The accuracy of phishing detection schemes is normally evaluated using a set of benchmark datasets.

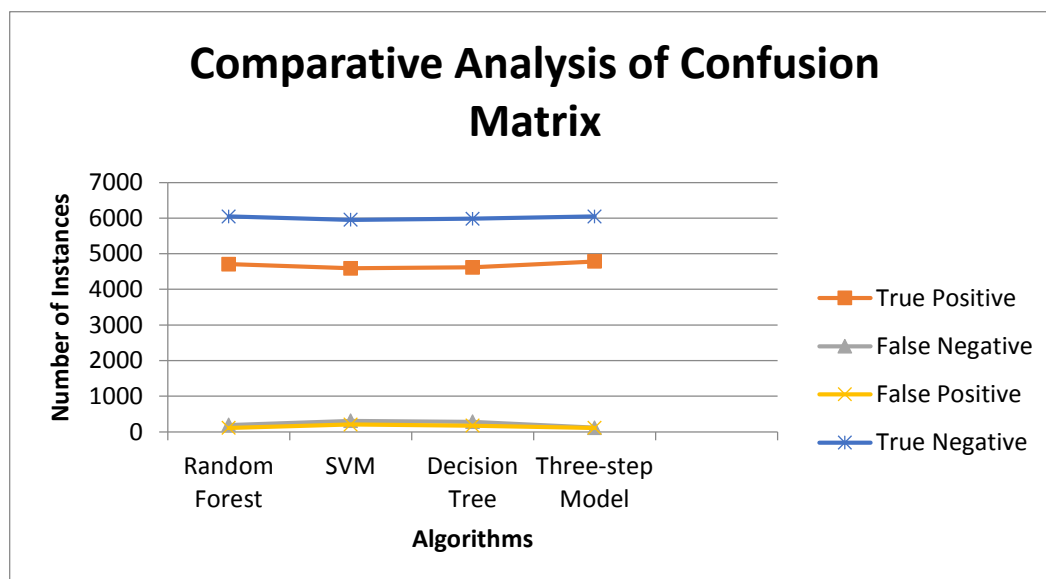


Figure 4.9: Weighted Average of Confusion Metric Comparison Among Learning Models.

A column standardized line synopsis shows the rates of accurately and erroneously characterized perceptions for each actual class. A section standardized segment outline shows the rates of accurately and inaccurately ordered perceptions for each predicted class.

The TP rate indicates the proportion of the number of phishing websites correctly labeled as phishing. An FP rate shows the proportion of the number of legitimate websites incorrectly labeled as phishing. A TN rate represents the proportion of the number of legitimate websites correctly labeled as legitimate, whereas the FN rate shows the proportion of the number of phishing websites incorrectly labeled as legitimate. Figure 4.9 is shown the Weighted Average of Confusion Metric. Weighted Average is how much your result deviates from the predicted and true values. The proposed TP size is the highest value 4782 better than other compared models. The confusion matrix shows the all-out number of perceptions in every cell. The lines of this compare to the actual class, and the sections relate to the predicted class. Corner to corner and off-slanting cells compare too effectively and inaccurately ordered perceptions, individually.

Chapter Five

Conclusion and Future Work

Finally, this chapter summarizes the whole thesis, demonstrates how the stated aims and objectives have been achieved, and proposes some areas for further study in the future.

5.1 Conclusion

In this research, we investigate the problem of website phishing using three combined detectors which are Random Forest, SVM, and Decision Tree (individually). These three detectors slightly varied in their results, yet all of them scored less accuracy than the combined ensemble to seek its applicability to the phishing problem. The proposed is implemented and evaluated using dataset. SVM multi-class classifier is used in this thesis for classification purpose. The experimental results over ARFF dataset showed that the accuracy enhancement of proposed compared to the detector is around 1.2. The three-step model is 1.2 % higher in accuracy than the Random Forest detector. It is 3.0996% more accurate than using the SVM alone. And finally, it is 2.584% higher in accuracy of detecting phishing websites than using Decision Tree individually. Thus, it is significantly effective in detecting phishing websites. As shown by the results of each detector which are Random Forest, SVM, and Decision Tree which scored detecting accuracies of 97.25%, 95.35%, and 95.87% respectively. Yet the ensemble scored higher value of accuracy than the highest among them which is (98.52%). It can be safely concluded that the ensemble proved its validity by benefiting from the variety of the three detectors. Furthermore, we demonstrated the shortcoming of using URL features such as URL lengths that seem to give higher accuracy but may not do so soon. Our feature extraction and classification times are very low and show that our approach is suitable for real-time deployment. Our approach is likely to be very effective in modern day phishing strategies like extreme phishing that are designed to deceive even experienced users.

5.2 Future Work

In future, we wish to explore the robustness of machine learning algorithms for phishing detection in the presence of newer phishing attacks. We are also developing a real-time browser

add-on that will provide warnings when visiting suspicious sites. The authors believe that the phishing attacks are increasing day by day based on the literature review, though ample solutions are available. However, it is a bit challenges to educate\trained the users besides of detecting phishing attacks.

References

Abdelhamid, N., Ayesha, A., and Thabtah, F. (2014). Phishing detection based Associative classification data mining. *41(13)*, 5948-5959.

Abdelhamid, N., Thabtah, F. and Abdel-Jaber, H. (2017). Phishing detection: A recent intelligent machine learning comparison based on models content and features. *In IEEE International Conference on Intelligence and Security Informatics (ISI)*, 22–77, China:IEEE.

Aburrous, M., Hossain, M., Dahal, K., Thabtah, F. (2010). Experimental case studies for investigating e-banking phishing techniques and attack strategies. *cognitive computation*, 2(3),242-253 .

Aburrous, M., Mohammed, R., Dahal, K., and Thabtah, F. (2011). Phishing website detection using intelligent data mining techniques. *designand development of an intelligent association classification mining fuzzy based scheme for phishing website detection with an emphasis on E-banking*, University of Bradford.

Aburrous, M., Hossain, M. A., Thabatah, F. and Dahal, K. P. (2008). Intelligent phishing website detection system using fuzzy techniques. In: *Proceedings of the 3rd International Conference on Information & Communication Technologies: From Theory to Applications (ICCTA'08)*. New York: IEEE.

Al-diabat, M. (2016). Detection and Prediction of Phishing Websites using Classification Mining Techniques. *International Journal of Computer Applications*, 147(5) .

Ayesha, S., Mustafa, T., Sattar, A., and Khan, M., (2010). Data Mining Model for Higher Education System. *European Journal of Scientific Research*, 43(1), 24-29.

APWG (2017) Global phishing survey: domain name use and trends in 2016.[online] Retrieved 12 Sep 2019, from <https://apwg.org/apwg-news-center/>.

Bahnse, A., Behrouz, E., Villegas, S. , Vargas,J., and González,F. ,(2017).Classifying phishing URLs using recurrent neural networks. *in Proc.IEEE APWG Symp. Electron. Res. (eCrime)*,1–8.

Ding,Y., Luktarhan,N., Li, K.,and Slamuw. (2019). A keyword-based combination approach for detecting phishing webpages,Computers & Security.

Fazliya, M.H F. ; Naleer, H.M.M. 2019. “A Rule Based Prediction of Phishing Websites Using Data Mining Classification Techniques.” *Journal of Technology and Value Addition* 1 (2): 31–40.

Data sets Retrieved 20 Dec 2019, from UCI website <https://archive.ics.uci.edu/ml/datasets.php>.

Ezziyyani, M., Bahaj,M. , and Khoukhi, F. (2017).Advanced Information Technology, Services, and Systems.Proceedings of the International Conference on Advanced Information Technology, Services and Systems,Springer.

Feng,F., Zhou,Q., Shen, Z. Yang,X., Han,L., and Wang ,J. (2018). The application of a novel neural network in the detection of phishing websites .*Journal of Ambient Intelligence and Humanized Computing* .

Gail,M.H.,Benichou ,J.,Armitage,P ., and Colton,T. (2000).Encyclopedia of epidemiologic methods.Publisher *John Wiley and Sons*, first edition,ISBN: 978-0-471-86641-1.

Hadia,W., Aburuba,F.,and Alhawarib,S.(2016). A new fast associative classification algorithm for detecting phishingwebsites *Applied Soft Computing*, volume 48, 729–34. Elsevier Science Publishers B.V.

Hyndman, R. J., and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22 (4): 679–688.

Huh J.H., Kim H. (2012) Phishing Detection with Popular Search Engines: Simple and Effective. In: Garcia-Alfaro J., Lafourcade P. (eds) *Foundations and Practice of Security*. FPS 2011. *Lecture Notes in Computer Science*, vol 6888. Springer, Berlin, Heidelberg.

Kulkarni, A., Brown. (2019). Phishing websites detection using machine learning. *International Journal of Advanced Computer Science and Applications (ijacsa)*, 10(7).

Luke ,I. (2020).The 5 most common types of phishing attack.[online] Retrieved 1 March 2020, from <https://www.itgovernance.eu/blog/en/the-5-most-common-types-of-phishing-attack>.

Mahalakshmi, A., goud,N.S.,and murthy ,G.V. (2018).A survey on phishing and it's detection techniques based on support vector method and software defined networking .*international journal of engineering and advanced technology (ijeat)*.8(2) , 2249 – 8958.

Mande, S., and Thosar,D.S. (2018).Detection of phishing web sites based on extreme machine learning. *International Journal of Advance Research And Innovative Ideas In Education Publisher (IJARIIE)*, 4 (6) ,111-114.

McHugh, and Mary L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*. 22 (3): 276–282.

Ming ,Q. ,and Chaobo ,Y., (2006).Research and Design of Phishing Alarm System at Client Terminal, APSCC'06, Proceedings of the 2006 IEEE Asia-Pacific Conference on Services Computing.

- Mohammad, M., Thabtah, F., McCluskey, L. (2014). Predicting Phishing Websites Based on Self-Structuring Neural Network. Intelligent rule based phishing websites classification. *Neural Computing and Applications*, 25(2), 443-458.
- Nagaraj, K., Bhattacharjee, B., Sridhar, A. and Sharvani, G.S. (2018). Detection of phishing websites using a novel twofold ensemble model. *Journal of Systems and Information Technology*, 20(3), 1328-7265.
- Nakashima, E., and Harris, Sh. (2018). How the Russians hacked the DNC and passed its emails to WikiLeaks. The Washington Post. Retrieved February 22, 2020.
- Nandhini, S., Vasanthi, V. (2017). Extraction of features and classification on phishing websites using web mining techniques. *International Journal of Engineering Development and Research (IJEDR)*, 5(4), ISSN: 2321-9939.
- Nazreen Banu, M., Munawara Banu, S. (2013). A Comprehensive Study of Phishing Attacks. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 4 (6), 783-786. ISSN: 0975-9646.
- Pandey, P. K., Singh, S. K. (2019). Phishing diagnosis: a multi-feature decision tree-based method. *International Journal of Engineering and Advanced Technology (IJEAT)*, 9(2), ISSN: 2249 – 8958.
- Patil, P., and Devale, P. R. (2017). A literature survey of phishing attack technique. *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE)*, 5(4), 198-200.
- Preethi, V., Velmayil, G. (2016). Automated phishing website detection using URL features and machine learning technique. *International Journal of Engineering and Techniques*, 2(5), 107–115. Retrieved 1 Dec 2019, from <http://www.ijetjournal.org>.
- Robert, P., and Marco, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*. 32 (15): 4407–4429.

Qabajeha, I., Thabtah,F.,and Chiclanaa,F.(2018). A recent review of conventional vs. automated cybersecurity anti-phishing techniques.Computer Science Review. Retrieved 20 Dec 2019, from <http://www.cse.dmu.ac.uk/~chiclana/publications/Computer-Science-Review-2018.pdf>.

Rathod, P.D., Kapse,S.R. (2017).Secure bank transaction using data hiding mechanisms. *International Conference on Innovations in Information Embedded and Communication Systems (ICIIECS)*. 6 (9), Coimbatore, India,IEEE. Retrieved 20 Dec 2019, from <https://doi.org/10.15680/IJRSET.2017.0609194>.

Seker R., (2006). Protecting Users against Phishing Attacks with AntiPhish.Journal Computer Software and Applications, 13(8), pp. 517-524.

Sharma, A., Singh,P., Kaur,A. (2016).Phishing websites detection using back propagation algorithm: a review.*The International Journal Of Engineering And Science (IJES)*. 5(5), 103-106.

Shetty,A.D.,and Chiplunkar,N.N. (2016).Anti-Phishing detection system to detect and prevent deceptive phishing in SoNet sites.*International Journal of Innovative Research in Computer and Communication Engineering(IJIRCCE)*. 4(5), 9204-9208.

Shrivasa, A. K., and Suryawanshi, R. (2017). Decision tree classifier for classification of phishing website with Info Gain feature. *Int. J. for Res. Appl. Sci. Eng. Technol.*, 5(5), 780–783.

Thabtah,F., and Kamalov,F. (2017).Phishing detection: a case analysis on classifiers with rules using machine learning.*Journal of Information & Knowledge Management* , 16(4): 1750034.

Ubing, Alyssa Anne, Syukrina Kamilia Binti Jasmi, Azween Abdullah, N Z Jhanjhi, and Mahadevan Supramaniam. (2019). Phishing website detection: an improved

accuracy through feature selection and ensemble learning. *International Journal of Advanced Computer Science and Applications. (IJACSA)*, 10(1).

Varshney,G., Misra,M.,and Atrey,P. (2016).A survey and classification of web phishing detection schemes.*Security and Communication Networks*,9(6),6266-6284. in Wiley [Online]Library (wileyonlinelibrary.com).

Willmott, C. J.,and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*. 30: 79–82.

Zhou, Zhi-Hua. (2012). *Ensemble methods: foundations and algorithms*. Chapman and Hall. Retrieved 2 Dec 2019, from <https://analyticsindiamag.com>.