

**AN EFFECTIVE APPROACH FOR PHISHING DETECTION USING MACHINE
LEARNING ALGORITHM**

BY

**UZZAL ROY
ID: 191-15-12102**

AND

**SAMIA ISLAM SUMI
ID: 191-15-12313**

This Report Presented in Partial Fulfillment of the Requirements for the
Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Fahad Faisal
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

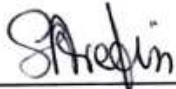
APPROVAL

This Project titled “**An effective approach for phishing detection using machine learning algorithm**”, submitted by Uzzal Roy, ID No: 191-15-12102, and Samia Islam Sumi, ID No: 191-15-12313 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 24-01-2023.

BOARD OF EXAMINERS

Dr. Touhid Bhuiyan
Professor and Head

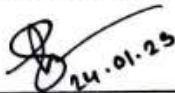
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

 24.01.23

Chairman

Dr. Mohammad Shamsul Arefin
Professor


Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

 24.01.23

Internal Examiner

Md. Sabab Zulfiker
Senior Lecturer

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

 24.1.2023

Internal Examiner

Dr. Ahmed Wasif Reza
Associate Professor

Department of Computer Science and Engineering
East West University

External Examiner

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Md. Sadekur Rahman, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

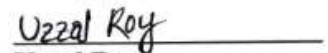


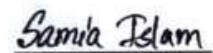
Md. Sadekur Rahman
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised by:


Fahad Faisal
Assistant Professor
Department of CSE
Daffodil International University

Submitted by:


Uzzal Roy
ID: 191-15-12102
Department of CSE
Daffodil International University


Samia Islam Sumi
ID: 191-15-12313
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Md. Sadekur Rahman, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “Machine Learning” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Dr. Touhid Bhuiyan, Professor, and Head**, Department of CSE, for his kind help to finish our project and also to other faculty members and the staff of CSE department of Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

Finally, we must acknowledge with due respect the constant support and patients of our parents.

ABSTRACT

The trend of tricking web users has kept pace with the expanding utilization of online surfing. The rise of phishing attacks postures a noteworthy risk to individuals and organizations everywhere. Phishing is constantly advancing to receive modern methods and techniques to steal important pieces of information from users. Phishing is a form of attack initiated by an email or social media message which mainly forwards the casualties to malicious web pages and these are extremely difficult to identify for security administrators. Phishing is a part of social engineering. Through this, hackers design a web page duplicate and send it to the user when the user enters information that data is directly saved to a database created by hackers. The most commonly used phishing techniques are link manipulation, filter evasion, website forgery, social engineering, and covert redirect. To recognize unique patterns, Machine Learning algorithms continuously learn from huge bulk data and in most research, it has been claimed that machine learning-based methods are more effective than other methods. Here, we use Five machine-learning classification techniques to detect phishing web pages and legitimate web pages with desirable accuracy. In our work, we apply Logistic regression, Decision tree, XGBoost, Random Forest, and SVM algorithms. All algorithms perform incredibly well on dataset. The Random Forest algorithm surpasses them all with a 98% accuracy rate.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	ii
Acknowledgments	iii
Abstract	iv
CHAPTER	
CHAPTER 1: INTRODUCTION	1-6
1.1 Introduction	1-2
1.2 Motivation	3
1.3 Rationale of the Study	3-4
1.4 Research Question	4
1.5 Expected Output	4-5
1.6 Report Layout	5-6
CHAPTER 2: BACKGROUND	7-14
2.1 Terminologies	7
2.2 Related Works	7-12
2.3 Comparative Analysis and Summary	12-13
2.4 Scope of the Problem	13-14
2.5 Challenges	14
CHAPTER 3: RESEARCH METHODOLOGY	15-20
3.1 Research Subject and Instrumentation	15
3.2 Data Collection Procedure	15
3.3 Statistical Analysis	16

3.4 Proposed Methodology	16-20
3.5 Implementation Requirements	20
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	21-28
4.1 Experimental Setup	21
4.2 Experimental Results & Analysis	21-27
4.3 Discussion	27-28
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	29-31
5.1 Impact on Society	29
5.2 Impact on Environment	30
5.3 Ethical Aspects	30-31
5.4 Sustainability Plan	31
CHAPTER 6: SUMMARY, CONCLUSION, RECOMMENDATION AND IMPLICATION FOR FUTURE RESEARCH	32-34
6.1 Summary of the Study	32
6.2 Conclusions	33
6.3 Implication for Further Study	33-34
REFERENCES	35-37

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Percentage of levels column	16
Figure 3.2: Proposed Methodology	17
Figure 3.3: Logistic Regression	18
Figure 3.4: Decision Tree	18
Figure 3.5: XGBoost	19
Figure 3.6: Random Forest	19
Figure 3.7: Support Vector Machine (SVM)	20
Figure 4.1: Heat Map for 0 to 10 columns	22
Figure 4.2: Heat Map for 10 to 20 columns	22
Figure 4.3: Heat Map for 20 to 30 columns	23
Figure 4.4: Heat Map for 30 to 40 columns	23
Figure 4.5: Heat Map for 40 to 50 columns	24
Figure 4.6: Confusion Matrix of LR	25
Figure 4.7: Confusion Matrix of DT	25
Figure 4.8: Confusion Matrix of XGBoost	25
Figure 4.9: Confusion Matrix of RF	25
Figure 4.10: Confusion Matrix of SVM	25

LIST OF TABLES

TABLES	PAGE NO
Table 2.1: Comparison between previous work	13
Table 4.1: Model Accuracy	21
Table 4.2: Classification Report	26

CHAPTER 1

Introduction

1.1 Introduction:

When hostile actors send messages posing as reliable people or organizations, this is known as phishing, it is a type of cybersecurity attack. Phishing messages mislead users into taking activities like downloading a malicious file, clicking on a risky link, or revealing private information like login passwords, confidential documents, etc. Attacks involving phishing and other forms of social engineering [28] are frequently combined with network attacks, malware, and other dangers like code injection. If the attacks are successful and the spammers get into the system of the organization or individuals then the victims will face a huge loss. Even an unstable phishing attack can cause billions of dollars of loss and those are very effective if we consider the last few years' records. Organizations and individuals will face huge losses if they come under a phishing attack. So, it is very important to identify those malicious web pages or emails through which the intruder is trying to get into devices or compromising information for their own motive.

Every year, more companies and people become victims of phishing scams and other dangerous online threats. The Anti Phishing Working Group recorded 1,270,883 phishing attacks between July and September of 2022, setting a new record and making them the worst attacks the APWG has ever encountered [27]. Most of these incidents are caused by human error and a lack of security measures. Even though an address appears to be legitimate does not necessarily imply that it is. Links and URLs that lead to malicious websites may appear to be nearly identical to those of trustworthy websites, but they really use different spelling, more special characters, or even a new domain. It's quite difficult for random users as well as for experts to identify those kinds of malicious sites with bare eyes. So, organizations and individuals should be trained or aware of this kind of malicious attack and needs to be updated to keep pace with modern technologies. The more worlds keep introducing modern technologies the chance of getting hacked by attackers also keeps increasing. In this modern world, it is very important to keep yourself safe from attackers by taking necessary precautions.

One of the most incredible methods to protect yourself is to use caution when browsing the internet and when encountering questionable links, as well as to understand how to detect a phishing website. Staying up to date on cybersecurity, continuously educating oneself, and being aware of the latest dangers and attacker techniques are critical. In many cases, the phishers do not directly cause economic damages but they violate the law and resell the information gathered from phishing in a secondary market for making money. They also take it as a challenge to do something new and for this, they try new techniques to enter into the device through malicious code. There are several anti-phishing methods available, including machine learning, black-and-white listing, virtual similarity identification, and heuristic detection. Various machine-learning algorithms for recognizing phishing scams on websites will be compared in this study. Nowadays to keep data safe it is very important for all of us to know about these things and get trained so if anything bad happens then we will be able to take immediate actions to minimize the damage.

In this paper, the features of phishing are extracted from the phishing web pages in this study to create an effective method for phishing detection. These features are then utilized to generate the feature vector required by the XGBoost classifier to train the proposed technique. Extensive studies show that the suggested anti-phishing strategy performs competitively on real-world datasets in terms of several assessment statistics. Logistic regression is a supervised learning model trained and tested dataset to show the high detection accuracy rate to predict the URL. Decision tree classifies with optimal feature selection for phishing website detection. According to other algorithms, Random Forest runtimes are relatively fast, and it can cope better with various websites and webpages for phishing detection, as well as SVM methods, which can handle both classification and regression on linear and non-linear data. All these algorithms perform quite well on our dataset. To detect phishing attacks by using the most effective machine learning algorithms is the main motive to do this work.

1.2 Motivation

It is an art for attackers to persuade their victims to trust misleading information. They fool people by creating exact web pages of the original web page to fulfill their own purpose. The attacker establishes a fictitious relationship with the victim to increase the likelihood that the victim will reveal private information to the attacker. People appear to be more receptive to friends or a familiar interface generally. Attackers are aware of this and are skilled at establishing this relationship, as well as being very misleading in posing as someone familiar. Furthermore, people may be willing to comply with requests from platforms they enjoy using. The attackers are also aware of this and utilize it to obtain data. Then they send malicious code or messages through email, when the users open the mail and enter the link then their information will be saved into the attacker's database. Then they use these data for their own motive. They can use these for any bad purposes. These can lead to direct damage to organizations or companies if their confidential data are leaked, and their image in society or the market will be damaged. Lack of a proper security system and experts, the chance of leakage of information is huge. So, from this motivation, we decided to do research on phishing detection which can detect phishing webpages perfectly and will be able to secure the personal information of the victims from the attackers. Our suggested algorithm will be able to detect phishing web pages that cannot be identified by looking at them.

1.3 Rationale of the Study

In today's world, various kinds of tools and organizations are introduced to people to get them safe from cyber-attacks. But not every time this can help the people from being attacked because the hackers are being upgraded with a pace of digitalization and finding new ways to send malicious data to compromise the victim's system or data. Keeping these in consideration, we have decided to work on the topic which is based on phishing detection. Nowadays, it is a great issue for everyone who is using the internet based on their daily work because intruders set traps for users when they get caught in those traps all information will be shared with the hackers. Hackers can accomplish anything with the

use of this information. They can do an illegal activity with someone's identity. They can also misuse the devices like computers, smartphones, and tablets to cause damage or corrupt the system to gather information about the user, and steal data. Once the system is compromised then the attacker can access all the information available in the system. With the data, they can blackmail or ask for ransom money or they can sell that information in exchange for huge money. So, it is a great concern for everyone how to minimize these attacks or how we can keep ourselves safe while browsing the internet. That's why we have taken a dataset based on phishing detection by using some machine learning algorithms, we have proposed one of them that has the highest accuracy.

1.4 Research Question

While conducting study, some questions about this work arise. The following are the primary concerns of our work:

- How to collect and preprocess phishing data?
- How to extract features from a phishing dataset?
- Which ML algorithm will perform better to detect phishing web pages from the dataset?
- What is its future scope of it?

1.5 Expected Output

Our intention to publicly release journal articles on projects relevant to our study. A research paper provides the opportunity to do extensive analysis fast. Cybersecurity (CS) is the field of study area of our proposed study. Cyber security encompasses a wide range of disciplines. For our work, we have chosen phishing detection because it has grown into one of the most dangerous of all attacks. Here, we did phishing detection for web pages. The expected outcome is which technique is more accurate to detect phishing from web pages. Web page phishing is common nowadays, attackers create a web page that looks similar to the actual page. For this reason, normal users cannot identify that the web page is malicious, they click on the page and give details like email address, password, etc. The

attacker creates a personal database that stores all this information that is taken from the user. To find these kinds of phishing we used machine learning methods in this study to obtain the outcomes. Machine learning algorithms are best when it comes to predicting accuracy. For these, we must train our dataset. We attempted to determine the motive behind phishing and why it is becoming popular across attackers. We hope to publish articles on these topics following this study, and with the aid of machine learning techniques, we will be able to reliably identify phishing websites.

1.6 Report Layout

In total, there are six chapters. Every portion is explored from many perspectives, and each chapter has several parts that are covered in detail. This report paper contains the following information:

Chapter 1 Sections of this chapter are discussed in the following –

- 1.1 Discuss about Introduction part,
- 1.2 Discussing about the Motivations,
- 1.3 Rationale of the Study
- 1.4 Discuss the Rational Study of this article,
- 1.5 Expected Outcome
- 1.6 Report Layout

Chapter 2 We discussed about,

- 2.1 Preliminaries/Terminologies
- 2.2 Related Works
- 2.3 Comparative Analysis and Summary
- 2.4 Scope of the Problem
- 2.5 Challenges

Chapter 3 In this chapter we described the whole working process of our work together with some sections,

- 3.1 Research Subject and Instrumentation
- 3.2 Data Collection Procedure/Dataset Utilized

3.3 Statistical Analysis

3.4 Proposed Methodology/Applied Mechanism

3.5 Implementation Requirements

Chapter 4 Experiment and Result Discussion of this research have discussed in this chapter

4.1 Experimental Setup

4.2 Experimental Results & Analysis

4.3 Discussion

Chapter 5 We talk about social impact in our society in the following chapter

5.1 Impact on Society

5.2 Impact on Environment

5.3 Ethical Aspects

5.4 Sustainability Plan

Chapter 6 We considered about,

6.1 Summary of the Study

6.2 Conclusions

6.3 Implication for Further Study

CHAPTER 2

Background

2.1 Terminologies

Firstly, we collected some allied papers to generate a literature review and tried to know about the phishing detection process. Because without knowing phishing detection methods, we cannot protect this. So, we read all these related papers carefully to gain knowledge about phishing. We studied some earlier work in order to implement our work flawlessly and to become familiar with this new term. Then we think we will be able to propose our work by using some machine learning approaches. Three types of ML are there: supervised, semi-supervised, and unsupervised. Here, we will use supervised machine-learning approaches to detect phishing sites. From the starting day, we research phishing detection. We read so many papers to gain knowledge about phishing. It was not a simple task for us to know phishing perfectly. Because there are lots of ways to attack through phishing. Phishing is totally a new form for people and us to know perfectly how it works.

We collect lots of papers related to this work then we read those papers very carefully so that we did not miss any things that are very useful for our work. After collecting papers and reading all the papers we start to know the machine learning algorithm. Because there are lots of machine learning algorithms. We research and read all algorithms because we need to find out the best algorithm for our proposed work. When we read all the related papers we discover so many new terms that are so useful for our work. We select five types of supervised machine learning algorithms for our desired dataset.

2.2 Related Works

In 2019, Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri proposed a method to detect phishing from URLs by using a machine-learning approach, they collected 36,400 legitimate URLs and 37,175 phishing URLs. In order to improve accuracy, they also divided their feature list into two distinct classes: word vectors and

features selected by humans. In this paper seven ML algorithms like DT, Adaboost, K-star, kNN, RF, SMO, and NB, and NLP-based features, word vectors, and hybrid features were used. Random Forest Model outperforms with an accuracy of 97.98%. [1]

In 2020, a paper was published about “Phishing Detection Using Machine Learning Technique”. [2] Authors mainly followed 3 steps for phishing detection. Firstly, they select URLs such as URL legitimate and URL phishing. Then they extract feature vectors from the URL by using vocabulary, host, and word. And lastly, they use 3 types of Machine learning algorithms like FACA, RF, and SVM. SVM shows the highest accuracy of 95.66%.

Authors of [3] attempt a method for anti-fishing approaches using a machine learning algorithm. They suggested analyzing the models based on their contents and features. The authors collect 11000 examples for the dataset and most of them are either binary or multi-values. These websites are from Phishtank and other sources. They use a variety of ML techniques, but eDRI and decision tree algorithms provided the most accurate predictive models. Moreover, Bayes Net and SVM showed good performance concerning the accuracy, but users having trouble understanding this model is its greatest issue.

In 2019, Mohammad Mehdi Yadollahi, Farzaneh Shoeleh, Elham Serkani, Afsaneh Madani, and Hossein Gharaee published a method for Phishing Detection Using Hybrid Features based on a machine learning approach. Here, they identify between legitimate and phishing websites. Their main aim is to detect malicious websites from URL. For the purpose a phishing detection system utilizing XCS, they have gathered about 3983 phishing websites and about 4021 safe websites. Additionally, they produce a comparison of their approach's effectiveness with a wide range of learning algorithms. Different types of algorithms are used here like DT, Adaboost, Kstar, RF, SMO, NB, and XCS. XCS gives the highest accuracy which is 98.3%. [4]

To classify emails as legitimate or phishing emails they have utilized statistical classification methods on websites. They also presented two unique features created by adaptive Dynamic Markov Chains and by latent Class-Topic Models. [5]

Authors of [6] have proposed content-based phishing detection using machine learning. Their dataset was collected from Phistank.com. Through the Python programming language, they were able to get the HTML content of relevant but nonexistent web pages. They use 8 different machine learning algorithms. They choose the algorithm very carefully and focused on analyzing the source codes and contents of websites and e-mails. By using these algorithms, they identified 58 different features. Here XGBoost algorithm has accuracy of 98% for that dataset. In the future, the authors wanted to use some hybrid models and deep learning models to increase efficiency of the system.

In [7] an effective method for phishing detection on Twitter has suggested. They use machine-learning approaches and they named their technique “PhishAri”. They use Twitter-specific features and URL features for phishing detection. Their data collection involves two steps. Firstly, they collect data from Twitter through API and then label the tweets as phishing or legitimate. For labeling tweets, they used two blacklists, PhishTank and Google Safebrowsing. They divided the features into four features such as URL based, WHOIS-based, Tweet Based, and Network-Based for phishing detection. There are two sections in their study. They establish a categorization model based on several features in the first section before deploying a Google Chrome extension to develop an end-user solution. In contrast to 1,473 tweets with unique text, they have 1,589 phishing tweets. They have the highest accuracy of 92.52% for the Random Forest Algorithm.

Authors in [8] introduced a method for phishing detection using an adaptive boosting approach. In this research paper, To determine the strongly associated features, the authors use a heat map and a features package from MATLAB. Because it is adaptable and simple, the authors here choose the AdaBoost classifier for identifying website phishing. They compile datasets made up of 30 features from the PhishTank archive, the MillerSmiles archive, and Google searching operators. They divided the features into four categories to

detect phishing sites. By using some Machine learning algorithms, they got the best performance with training percentage of 70% and accuracy and F-measure are approximately 99%.

The authors [9] described a proposal for phishing detection using a URL-Based Heuristic. They have proposed a new technique where they have 11,660 phishing sites and 5,000 legitimate sites. Here the authors design a system model where they have 6 phases. Every phase has different types of processes like selecting features, calculating six values of the heuristics, etc. In the last phase, they Classify the websites by using some machine learning approaches. And the ML technique can identify over 97% of phishing sites.

A URL-based detection system, combining the URL of the web page URL and the URL of the web page source code as features [10] here is a system designed for unknown phishing pages which provide high accuracy and low false positive rate detection results.

The authors [11] use a data set of 2889 phishing and some emails for predicting phishing emails. Here they use six types of classifiers to get accurate accuracy.

In this study, [12] they provide a framework for intelligent phishing website identification that employs various machine learning methods to distinguish between legitimate and phishing websites. The experiment shows that Adaboost with SVM shows the best accuracy of 97.61%.

The authors [13] proposed the RRFST method to detect phishing emails. They use some other techniques like DT, CART, and FST. But FST gives a great accuracy of 99.27%.

In this research, a Random Forest algorithm-based model for categorizing and detecting phishing sites based on 63 features has been developed. The proposed model has a high accuracy of 96.91% and a low error rate of 0.03%. [14]

Nuttapong Sanglerdsinlapachai and Arnon Rungsawang proposed a method to detect web phishing. Firstly, they use 200 web data that consist of 100 phishing and 100 non-phishing. After using some ML approaches their accuracy boosted to 92% for f-measure. [15]

In this paper, the authors [16] take a total of 10000 pieces of data which contains 5000 legitimate and 5000 phishing. After implementing some algorithms, they get 20 features out of 48 and they get 98.11% best accuracy for Random Forest.

To detect spam emails, the authors [17] apply some machine learning algorithms to detect those fraudulent things easily. Their main aim is to detect spam emails but they have a limitation due to class conditional independence. In their proposed model, they get 98% accuracy for Naïve Bayes.

In this study [18], they apply the CNN model for detecting phishing sites. They take a total of 11055 data which shows 6157 genuine and 4898 phishing websites. They apply different types of CNN models and get higher accuracy of 96.6%. their phishing detection rate is 98.2% with a 97% F1-score.

Their main aim is to see the current impact of phishing strategies by using machine learning algorithms. They work on this to know how phishing URL system works on detection. SVM gives the best accuracy of 65.62% to 88.73% after four years. [19]

In this paper, phishing detection by using the CNN model has proposed. They take the all features by the website URL. They use Naïve Bayes, Logistic Regression, random forest, XGBoost, deep neural networks, recurrent neural networks, recurrent convolutional neural networks, etc. for their work and get an accuracy of 98.58%, 95.46%, and 95.22% on benchmark datasets. [20]

The authors [21] use five ML algorithms like Logistic Regression, Decision Tree, Naïve Bayes, KNN, and SVM to get correct information that a mail is a spam or not. Random forest and KNN give the best accuracy of 99%. They use the Weka tool for training and testing their dataset.

In this paper [22] they use the collected features for website phishing detection. They apply so many algorithms and Random Forest performs so well with an accuracy of 96%. After applying hybrid features the accuracy increased to 96.83%.

Improving the detection method of detecting phishing websites using machine learning technology is their main motive to do this work. They used the dataset in a 90:10 ratio as the train and test and gained 97.14% accuracy by the Random Forest algorithm with the lowest false positive rate. [23]

This experiment consists of two parts. One is to detect phishing via a newly registered domain and another is a proposed model. Their suggested Intelligent Phishing Detection (IPD) system is capable of actively addressing phishing detection issues. [24]

To protect the customers performing online transactions a new approach based on a Neuro-Fuzzy scheme to detect phishing websites is presented in this paper. Also claimed by adding more features and parameters in the future accuracy can be developed with a plugin toolbar for real-time applications. [25]

In this paper, They demonstrated how phishing URLs that contain a brand name tightly coupled with one or more phishing terms are undetectable by a classifier based on lexical features. To reduce these hindrances, different bag-of-X representations are explored by them including bag-of-words, segmented bag-of-words, and bag-of-n-grams. [26]

2.3 Comparative Analysis and Summary

After reviewing those papers, some similarities we found with them. When we read all the related papers we see that we have some new methods and some new terminologies that are very helpful for our work. At that time, we pick that method and researched it. After researching all those new terms, we finally realized that those new terms are so much useful for us. The authors of those papers describe and present their work very easily. They briefly describe those new terms and common terms so easily. They use lots of ML algorithms like random forest, decision tree, logistic regression, XCS, support vector machine, etc. They describe that algorithm very descriptive way so that the new researcher can easily understand this. In this work, we have used logistic regression, random forest, support vector machine, XGBoost, and decision tree algorithms. As a result, most of the work is so close to them. Most of the time they use the same machine learning algorithm that we

already use. That’s why we pick this type of paper so that we can compare it with our proposed model.

Table 2.1: Comparison between previous work

Title	Best Algorithm	Accuracy
Machine learning-based phishing detection from URLs	Random Forest	97.98%
Phishing Detection Using Machine Learning Technique	Support Vector Machine	95.66%
An Adaptive Machine Learning-Based Approach for Phishing Detection Using Hybrid Features	XCS	98.3%
PhishAri: Automatic Realtime Phishing Detection on Twitter	Random Forest	92.52%

According to the table, they use the Random Forest algorithm and it shows 97.38% accuracy. They use the XGBoost algorithm for detecting phishing and it gives 98% accuracy. For detecting phishing attacks using hybrid features, they employ the XCS Machine Learning-Based Approach and it shows 98.3% accuracy. They also use the SVM and it has 95% accuracy.

We can see that most of the time they use the Random Forest Algorithm but here XCS shows the best accuracy of 98.3%. Our work is “An effective approach for phishing detection using machine learning algorithm”, to detect those phishing sites where most of the authors faced difficulties.

2.4 Scope of the Problem

Phishing attacks are among the most prevalent types of attacks nowadays in the world. Every moment lots of people face this problem. Our target is to detect these phishing sites. We collect our dataset from Kaggle and we select some machine learning algorithms. Collecting any desired dataset was not easy for us but we tried a lot to find it. We are trying to apply that algorithm to our dataset to classify success rate prediction. In Kaggle there

are so many related papers on phishing detection. We read and observe all the papers and finalized a dataset for our work. Because most of the dataset is not well organized and not familiar to our work. After collecting the dataset, we face problems with algorithm selection. There are lots of algorithms and we need to select the right algorithm that will be so correct for our work.

2.5 Challenges

Our work is about detecting a phishing website by using a machine learning algorithm. We faced some difficulties while doing this proposal. Firstly, we faced a problem with finding a perfect dataset. We collect our dataset from Kaggle. Kaggle has lots of datasets on phishing detection. The tough task was to find the best dataset because most of the dataset was not up to mark or had already been implemented. Our main aim was to collect an appropriate dataset that will be so easy for us to implement.

After collecting our desired dataset, we faced another challenge and that is how should we solve this dataset. We think that we will apply the machine learning algorithm here. Then we select we will apply supervised approaches in our dataset. But another problem arises here because there are so many supervised algorithms. Some algorithms like SVM, RF, XGBoost, LR, and DT are implemented on our desired dataset. These five algorithms will give us a better result on our dataset. When we finalized our dataset, we are so confused about algorithm selection. We research a lot to find the best algorithms that will be so useful for our dataset. Then we select five supervised machine learning algorithms. When we started to apply algorithms to our dataset we face another problem. Our dataset is overfitting and we cannot use any overfitting data. Then we carefully see the dataset and finalized it to train our dataset first. When we moved on to train we see that our datasets all columns are not perfect. We need to rename some column's names. After that, we see that one column is not useful for us. We do not need the column for our work. Then we decide that we need to drop that column from our dataset.

CHAPTER 3

Research Methodology

3.1 Research Subject and Instrumentation

In this research, supervised machine learning algorithms are used because of having input and output data and we have to train a model in our work, we must employ Supervised Learning techniques for this. Supervised learning is good at classification and regression problems, the underlying patterns and connections between the input data and the output labels can be described by training the model, with never seen data enabling it to give accurate labeling results when presented. We used supervised classifier techniques in our work, and we applied five widely used algorithms to our dataset. They are logistic regression, decision tree, XGBoost, Random Forest, and SVM. After vectorizing data, it was divided into two portions: one is training data and other is testing data. It was separated into 80% and 20% groups. 80% data were kept for training and the rest for testing. We analyzed data for further implementation. We also extract our dataset by feature selection. We leveled our dataset into 0 and 1 which indicates phishing and non-phishing data on our dataset. To do our work we have used python and google Colab. Our used algorithms on the dataset are briefly discussed in the proposed methodology. We also have visualized a table that compares all the algorithm's accuracy we have used in this study.

3.2 Data Collection Procedure

Finding questions regarding any research-relevant dataset is a primary requirement. As our domain is cyber security related so data collection was quite difficult for us. There are some datasets available online. After searching for a few days we found a dataset that was available on Kaggle. In this work, we have applied and evaluated five different machine learning algorithms on our dataset containing 48 features extracted from 5000 phishing webpages and 5000 legitimate webpages. Here we have 10000 rows and 50 columns.

3.3 Statistical Analysis

Our dataset was acquired from Kaggle and saved it in CSV (Comma Separated Value) format to run in Google Colab. There are 50 columns and 10,000 rows available in our dataset among them 5000 data are leveled as a phishing class and 5000 are leveled as a non-phishing class. Here we used 8000 samples for training and 2000 samples for validation.

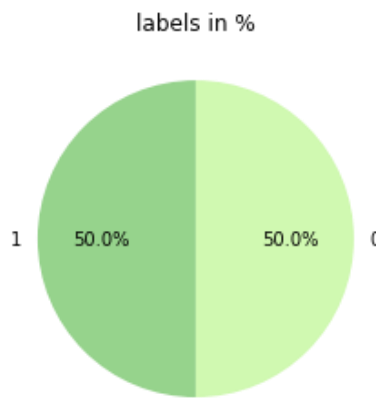


Figure 3.1: Percentage of levels column

3.4 Proposed Methodology

We have collected our dataset from Kaggle. To make the dataset machine acceptable for more accurate outcomes we have preprocessed our dataset. To do preprocessing we used python and imported some necessary libraries to do the task. After that, we have done Feature engineering to simplify the dataset and find out the more correlated features to obtain better accuracy. Then our data was converted, and it was split into training data and testing data. We separated our data into two groups: 80% and 20%. We reserve 80% of the data for training and the remaining 20% for testing. We used supervised classifier techniques in our work, and we applied five distinct algorithms to our dataset: logistic regression, decision tree, XGBoost, Random Forest, and SVM. All these algorithms

perform quite well on our dataset and Random Forest outperforms others on our dataset with 98% accuracy. Based on our phishing data, we will quickly outline the algorithms and their performance below.

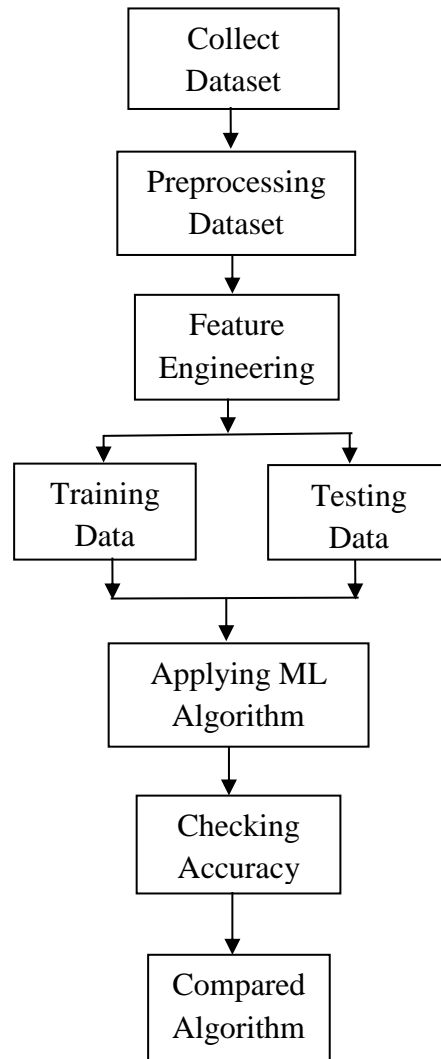


Figure 3.2: Proposed Methodology

3.4.1 Logistic Regression

A widely used statistical technique used to make binomial predictions referring to two classes is Logistic regression. Firstly, we will need data to feed our machine learning, then

the method will be easily interpreted and will give good results by classifying phishing and non-phishing data. The accuracy of Logistic regression was 93% for our dataset.

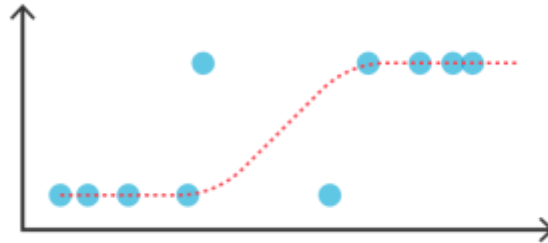


Figure 3.3: Logistic Regression [29]

3.4.2 Decision Tree

A decision tree employs a tree-like structure to determine the best option from various categories, which is often utilized in operations. After implementing the decision tree our dataset showed 94% accuracy.

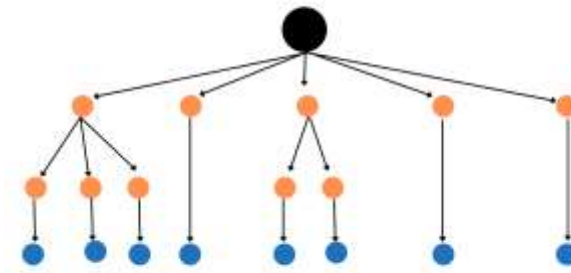


Figure 3.4: Decision Tree [30]

3.4.3 XGBoost

A regularizing gradient boosting framework is provided by the open-source software package known as XgBoost for the programming languages C++, Java, and Python. For creating infinitely better outcomes than other AI calculations XGBoost is notable. XGBoost has 97% accuracy rate for our dataset.

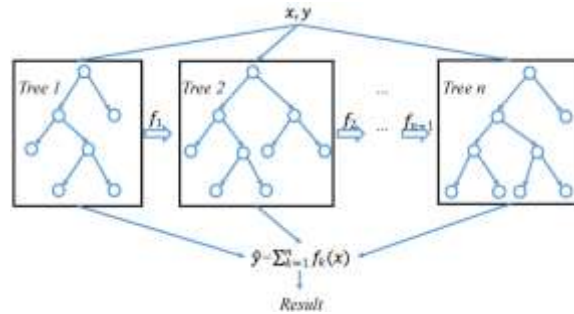


Figure 3.5: XGBoost [31]

3.4.4 Random Forest

A supervised learning algorithm is Random Forest algorithm that is very popular among researchers for classification and regression problems in ML. As random forest combines several decision trees, hence it is a long process, it gives better accuracy than other algorithms. If we consider other ML algorithms used in our dataset, then we will be able to see Random forest has the highest accuracy, which is 98% for our dataset.

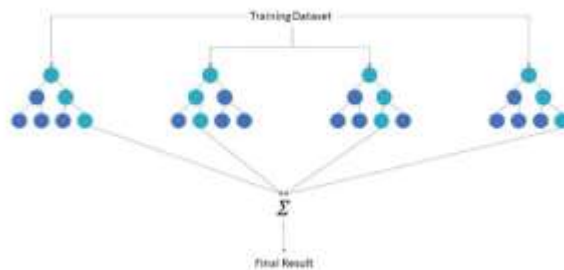


Figure 3.6: Random Forest [32]

3.4.5 Support Vector Machine (SVM)

SVM is used to solve classification and regression problems. Though, Machine Learning is basically utilized for Classification issues. Support Vector Machines are supervised algorithms that employ classification and regression analysis, which is generally used to separate data. After implementing the Support Vector Machine, our dataset showed 94% accuracy.

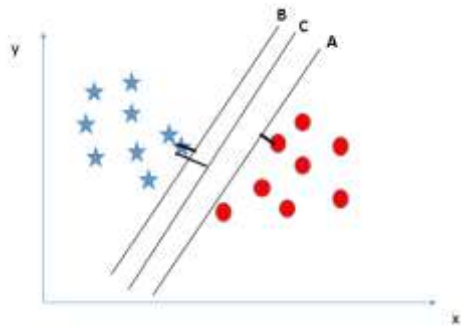


Figure 3.7: Support Vector Machine (SVM) [33]

3.5 Implementation Requirements

An Effective Approach for Phishing Detection Using Machine Learning Algorithms is the title of this research study. By using machine learning algorithms, we can detect phishing web pages is our main motive in this research. We collected our dataset from Kaggle and implemented five supervised algorithms to see which model gives us better or desired accuracy. To do this task, we will require a high-end laptop with a GPU and other specialized equipment. An inventory of the hardware and software required to launch our model is given below-

Software and Hardware:

- Intel(R) Core (TM) i3-10110U CPU @ 2.10GHz, 2592 Mhz, 2 Core(s)
- Physical Memory (RAM) 8.00 GB
- Google Colab

Tools use for Development:

- Microsoft Windows 11 Pro
- NumPy
- Pandas
- Seaborn
- Matplotlib

CHAPTER 4

Experimental Results and Discussion

4.1 Experimental Setup

We tried to set things in a new way to see which one is more effective and efficient for our experiment in this research. As we know no ML algorithm can give a 100% accuracy rate on any dataset. If it happens then we consider the dataset overfitting. At first, we also faced some problems while running our dataset and it was not making accurate predictions on testing data. After dropping the id, column we get balanced accuracy on every technique we used on the dataset. Our Logistic regression, Decision tree, XGBoost, Random Forest, and SVM model accuracy were 93%, 94%, 97%, 98%, and 94% respectively. Our model was also evaluated by some criteria including precision score, recall score, and f1- score.

Table 4.1: Model Accuracy

Name	Accuracy
Logistic Regression	93%
Decision Tree	94%
XGBoost	97%
Random Forest	98%
Support Vector Machine	94%

4.2 Experimental Results & Analysis

4.2.1 Heat Map

As our dataset has 10000 rows and 50 columns. It was quite difficult for us to show all features in one heatmap. So, we divided the dataset into 5 parts and now we can see the correlation clearly visible between the features.

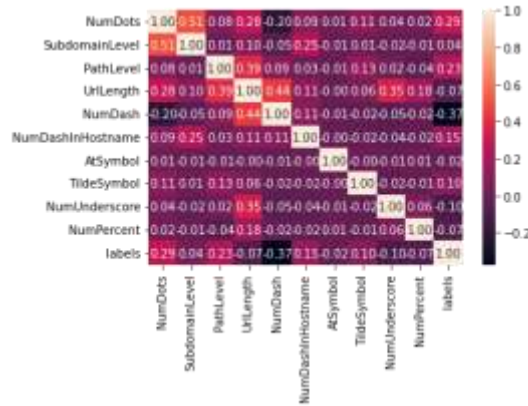


Figure 4.1: Heat Map for 0 to 10 columns

- At first, we took 10 columns against labels, and we can conclude that none of the features have a strong correlation with the labels. However, NumDash has some significant negative effects on the labels.

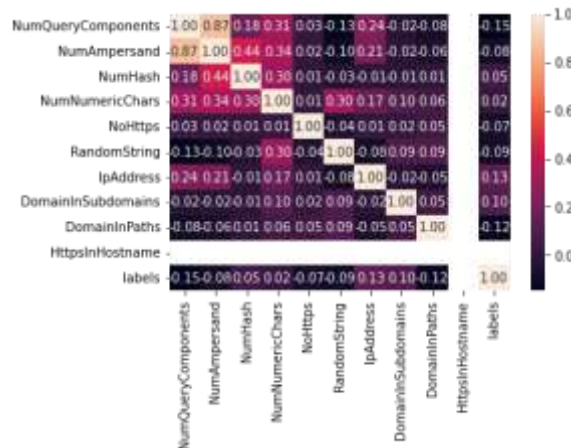


Figure 4.2: Heat Map for 10 to 20 columns

- Next we took the 10 columns and generated a heatmap for this. From the heatmap, we can see there are no strong or even medium-level strength correlation features with labels.

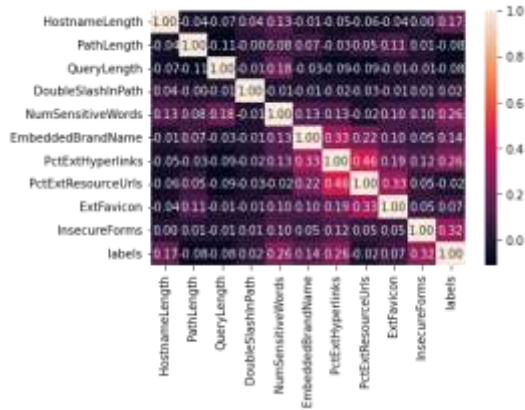


Figure 4.3: Heat Map for 20 to 30 columns

- For the third heatmap we have taken columns 20 to 30 to see the correlation between them. From the generated heatmap we can see there are no strong correlation features available in these columns.

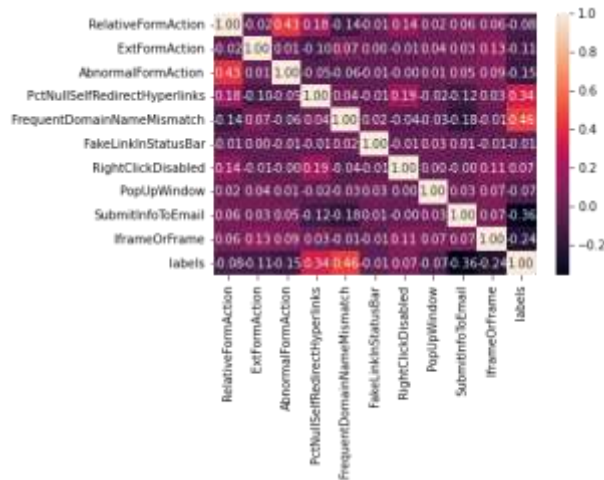


Figure 4.4: Heat Map for 30 to 40 columns

- Well in the fourth heatmap we have a few features that are linearly correlated to our variable. InsecureForms shows that as the value is higher so the probability of being a phishing site is higher. PctNullSelfRedirectHyperlinks shows the same positive correlation as InsecureForms. FrequentDomainNameMismatch shows that

it has medium linear correlation in positive direction. SubmitInfoToEmail seems to indicate that sites that ask users to submit their details to emails seems to be more high probability for phishing.

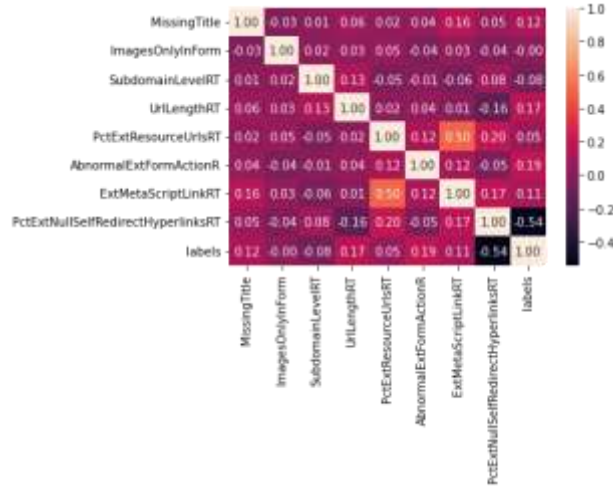


Figure 4.5: Heat Map for 40 to 50 columns

- The only column in this group that has some correlation with labels is PctExtNullSelfRedirectHyperlinksRT and it has a negative effect on labels which could mean that when the number of percent of null self-redirect hyperlinks occurs hence the probability of phishing increases.

4.2.2 Confusion Matrix

In a classification of a problem, the confusion matrix shows the summary of prediction outcomes. It interprets the result in a tabular form to understand the outcomes of the problem at a glance. An NxN matrix has two parts including actual value and predicted value. This matrix helps to understand whether any kind of errors are present or not in the dataset.

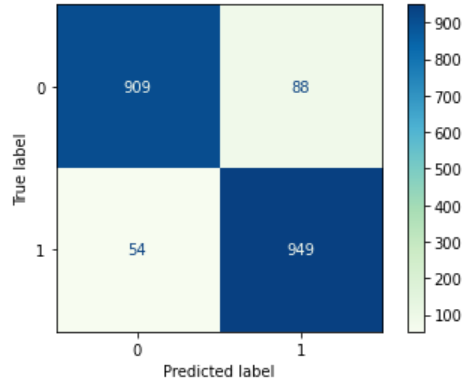


Figure 4.6: Confusion Matrix of LR

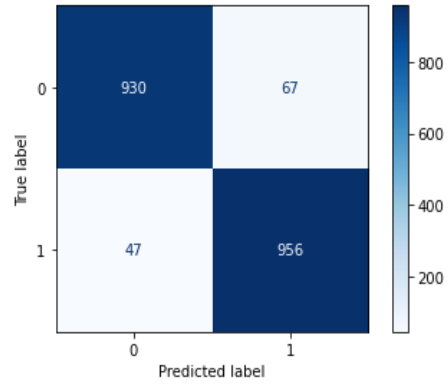


Figure 4.7: Confusion Matrix of DT

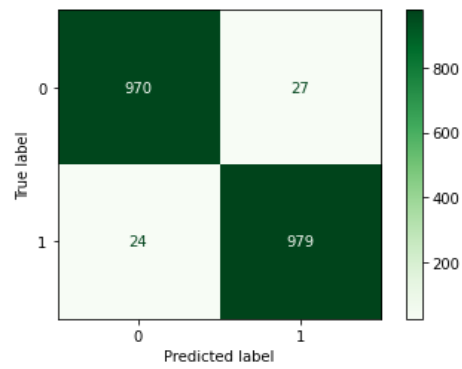


Figure 4.8: Confusion Matrix of XGBoost

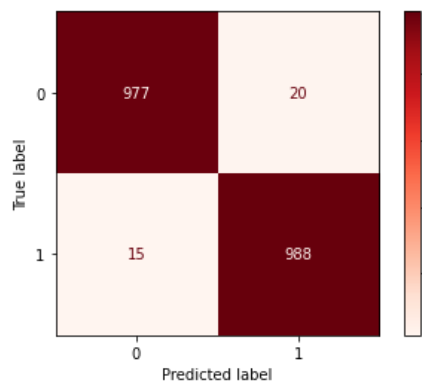


Figure 4.9: Confusion Matrix of RF

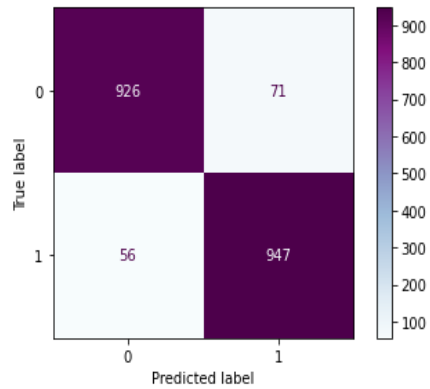


Figure 4.10: Confusion Matrix of SVM

Table 4.2: Classification Report

Algorithms	Level	Precision	Recall	F1 Score	Accuracy
Name		Score	Score		
Logistic Regression	0	0.94	0.91	0.93	0.93
	1	0.92	0.95	0.93	
Decision Tree	0	0.95	0.93	0.94	0.94
	1	0.93	0.95	0.94	
XGBoost	0	0.98	0.97	0.97	0.97
	1	0.97	0.98	0.97	
Random Forest	0	0.98	0.98	0.98	0.98
	1	0.98	0.99	0.98	
Support Vector Machine	0	0.94	0.93	0.94	0.94
	1	0.93	0.94	0.94	

The following table shows the comparison between the precision score, recall score, F1 score, and accuracy score. We have split our dataset into 0 and 1. One is for phishing and another is for non-phishing data.

In the logistic regression algorithm for level 0 precision score, recall score and F1 score are 0.94, 0.91, and 0.93 respectively. And for level 1 precision score, recall score and F1 score are 0.92, 0.95, and 0.93 respectively and the predicted accuracy for the logistic regression model is 93%.

In the Decision tree algorithm for level 0 precision score, recall score and F1 score are 0.95, 0.93, and 0.94 respectively. And for level 1 precision score, recall score and F1 score

are 0.93, 0.95, and 0.94 respectively and the predicted accuracy Decision tree model is 93%.

For the XGBoost algorithm, the level 0 precision score, recall score, and F1 score are 0.98, 0.97, and 0.97 respectively. And for level 1 precision score, recall score and F1 score are 0.97, 0.98, and 0.97 respectively and the predicted accuracy Decision tree model is 97%.

In the Random Forest algorithm for level 0 precision score, recall score and F1 score are 0.98, 0.98, and 0.98 respectively. And for level 1 precision score, recall score and F1 score are 0.98, 0.99, and 0.98 respectively and the predicted accuracy Decision tree model is 98%.

In the Support Vector Machine algorithm for level 0 precision score, recall score and F1 score are 0.94, 0.93, and 0.94 respectively. And for level 1 precision score, recall score and F1 score are 0.93, 0.94, and 0.94 respectively and the predicted accuracy Decision tree model is 94%.

4.3 Discussion

In this experiment, we evaluated the performance of various machine learning algorithms. We evaluated the performance of those algorithms to identify the best combination in detection. We also used performance measures like accuracy to show which algorithm works better in our dataset. Phishing causes harm or damage to all of us. Therefore, it is critical to identify phishing as soon as possible. As a result, the approach structure must be reasonable and authentic. It will be challenging to uncover phishing, as attackers use new techniques to get into the system and take important information from the user.

We have utilized Machine learning algorithms that can work on computational strategies. Also, it finds the common patterns in data that generate insight and makes a difference to create way better choices and predictions. The calculations adaptively move forward with their execution as the number of samples available for learning increases. The machine learning strategies have further been classified into supervised and unsupervised

approaches. In this study, we have used supervised learning algorithms to find which algorithm between logistic regression, decision tree, XGBoost, Random Forest, and SVM has the highest accuracy. Among these algorithms, Random Forest detects phishing more accurately with an accuracy of 98%.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

Our discoveries will have a profound impact on society. We mainly work on some websites and detect phishing there. Phishing is a technique of social engineering in which an attacker deceives a target into divulging sensitive information by sending them a message. If any person does not know what is Phishing and how it works he/she can face a great problem. In our society different types of people live here. Most people are educated here but not everyone does not know about this fraudulent work, especially phishing. Nowadays people use technology so much that the attacker can easily occur this type of crime. Here people are not aware of phishing and the attacker takes advantage. Just think of any organization where many people work for their daily income. If that person is a victim of the phishing attack then the person will be depressed and that will impact society very badly. Every person lives in a society. Here lots of types of people stay their life. Phishing is totally a new thing for them. Because all people are not educated and not so smart. Some people are educated but most of them do not know about this phishing attack. Some people do not hear this sound which means phishing word is new to them. Our country is a developing country. Nowadays people are using technology like smartphones, laptops, computers, etc. They also use Facebook and YouTube, and for some documentary reason, they use email. And this phishing attack occurred through this email. The attacker sends some links or messages via email. The attacker sends some unusual things through emails. When people open that email and that link then when they press those links that time their important records go to the attacker. The attacker easily gets all the secure information of the people. When people are affected by these attacks they get so depressed and create some unwanted things in their life. As a result, society is negatively impacted.

5.2 Impact on Environment

A phishing attack is a common attack for us because we do not aware of this type of attack. Considering that society, which includes a variety of human types, constitutes an environment. In an environment, different types of people stay here. If the entire people do not know about cybercrime like phishing, they can face trouble for their life. Because this phishing attack is a new word for people in any environment. If they know about phishing attacks they can easily overcome their mistake and can relieve themselves of their valuable life. That's why we work on this project so that we can give some ideas to the people and the environment can be free from danger. The environment is made up of society and society made up of people. In a society, there are lots of people living there. To make their lives easy and comfortable they use so many technologies. They use smartphones for their daily use. By using smartphones, they use emails in their daily life. Phishing attacks mainly occurred by email. With a simple tap, the attacker takes our important information and then uses it for many unusual works. Every people live in an environment. So, even if just one person is affected by this attack, it has a significant influence on the habitats.

5.3 Ethical Aspects

Our proposed work does not violet any people because we do not harm anyone and do not spread false news to them. Here we maintain ethics so much that we can give the right news to the people. We know that ethics is a moral philosophy. We already mentioned that we collect our data from Kaggle. Here we do not claim that the dataset is created by us.

Ethics makes a man perfect for their life. If any person does not follow ethics then they will not be able to lead their lives so comfortably. In our work, we find that we need to be ethical to our work. We collect our dataset from Kaggle then we apply some machine learning algorithms to this. Here we apply that algorithm by using Google Colab and this Google Colab is run on our own computer. We do not claim that we use another computer. To create our work, we do not use others' equipment and not stole any important code or data from others. To be ethical we do not copy any code that is available on the internet.

While doing our work we maintain our honesty and integrity. We have some ethical aspects:

- We do not copy anything from others,
- We do not use others' devices, equipment,
- We do not steal any secure information from others,
- We give our full strength to make our work perfect.

5.4 Sustainability Plan

A phishing attack is a new thing for us. We do not have proper knowledge about it. We do not know how it happens and what is the solution to it. Our main aim is to give proper knowledge to the people. Firstly, we collect a dataset then we apply different types of algorithms to it. The main reason why we generate this proposal is to give a perfect solution to the general people so that they do not face any phishing attacks. We create this model so that they can get correct information about phishing attacks and can spread this information to others people. If one person can get an advantage from our work then we will be so satisfied that we can do anything for the general people. Lastly, our simple work will sustain people from phishing attacks. When we started our work, we were so confused about our work. At that time, we did not know what we do. Then, we make a plan and started to execute that. Our main target was to create our work for the people who are so encouraged to improve our work.

CHAPTER 6

Summary, Conclusion, Recommendation and Implication for Future Research

6.1 Summary of the Study

Anti-phishing measures aim to prevent phishing assaults. Phishing is a type of cybercrime that is getting admired by hackers and attackers. They try to act like real web pages or email the victims. In this research, we focus on the real-time identification of phishing web pages by examining the URL of the web page using various machine learning methods (five of which are implemented and compared in the paper) as well as several feature sets. Not only the dataset but also the learning algorithm used here to wrench out features from the dataset was quite difficult. So, to complete our work have followed some steps they are as follows-

Step 1: Collected our desired dataset from Kaggle

Step 2: Preprocessing our dataset

Step 3: Feature Engineering

Step 4: Trained and tested our dataset

Step 5: Applied machine learning algorithms

Step 6: Accuracy checked on our dataset.

Step 7: Compared algorithms

After completing all the steps, we were able to detect phishing web pages. Among all the cyber security branches phishing is one of the most threatening attacks. The spammers make fools of the victims by acting like real and trustworthy sites. Users don't know or can't tell which sites are real and which are fake because the attackers create look-alike web pages like the authentic ones. As many online users are prey to attackers our work will have some impact on society and will be able to make users more aware of it.

6.2 Conclusions

In this paper, we conducted a comparison-based work for identifying phishing from web pages. As the ubiquity of using the internet is expanding quickly, the chance of attacks is additionally expanding, which has tremendous negative impacts on people, and society. It is a great challenge for researchers to detect these phishing sites as they are growing tremendously. Attackers tend to collect all the important and confidential data through a malicious attack. In this work, we have applied and evaluated five different machine learning algorithms on the dataset. Our dataset was collected from Kaggle containing 49 features which was extricated from 5000 phishing webpages and 5000 legitimate webpages. To increase the effectiveness of detection, machine learning algorithms are best. In this study among supervised and unsupervised machine learning algorithms, we have implemented supervised learning algorithms.

Five machine-learning classification techniques are implemented in our dataset to detect phishing from web pages and legitimate web pages with desirable accuracy. In our work we applied Logistic regression whose accuracy was 93%, the Decision tree has 94% accuracy, XGBoost has 97%, Random Forest has 98%, and the SVM algorithm has 94% Accuracy. All algorithms perform quite well on our dataset. Among them, the Random Forest algorithm came with the preferable accuracy which is 98%.

6.3 Implication for Further Study

There are some constraints and limitations in our dataset. As we all are aware of that, every model is designed for future improvement because An ongoing process, experimental research becomes better every day. Lack of awareness of people is one of the main reasons to be a victim of this type of attack. In the future, more options will be available for researchers as these kinds of phishing are increasing rapidly. New ways of attack will be introduced and it will give researchers new scope to do research. In this study, we are already aware that for our dataset Random Forest model is the best fit due to its high rate of accuracy. In the context of the future study of logistic regression, Adaboost models have

higher accuracy. These models can be used with different parameters for the model to get higher accuracy. As the dataset is available online by adding more features the study can be more precise and on a large scale. And also, a phishing detection site can be built, which can detect, recognize and gum up malicious websites without the participation of users.

References:

- [1] O. K. Sahingo, E. Buber, O. Demir, and B. Diri, "Machine learning based phishing detection from URLs," *Expert Syst. Appl.*, vol. 117, no. October, pp. 345–357, 2019, doi: 10.1016/j.eswa.2018.09.029.
- [2] J. Rashid, T. Mahmood, M. W. Nisar, and T. Nazir, "Phishing Detection Using Machine Learning Technique," *Proc. - 2020 1st Int. Conf. Smart Syst. Emerg. Technol. SMART-TECH 2020*, pp. 43–46, 2020, doi: 10.1109/SMART-TECH49988.2020.00026.
- [3] N. Abdelhamid and H. Abdel-jaber, "Learning Comparison based on Models Content and Features," *Ieee*, pp. 72–77, 2017.
- [4] M. M. Yadollahi, F. Shoeleh, E. Serkani, A. Madani, and H. Gharaee, "An Adaptive Machine Learning Based Approach for Phishing Detection Using Hybrid Features," *2019 5th Int. Conf. Web Res. ICWR 2019*, pp. 281–286, 2019, doi: 10.1109/ICWR.2019.8765265.
- [5] A. Bergholz, G. Paaß, F. Reichartz, S. Strobel, and J. H. Chang, "Improved phishing detection using model-based features," *5th Conf. Email Anti-Spam, CEAS 2008*, no. September, 2008.
- [6] U. Ozker and O. K. Sahingo, "Content Based Phishing Detection with Machine Learning," *2020 Int. Conf. Electr. Eng. ICEE 2020*, 2020, doi: 10.1109/ICEE49691.2020.9249892.
- [7] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on twitter," *eCrime Res. Summit, eCrime*, pp. 1–12, 2012, doi: 10.1109/eCrime.2012.6489521.
- [8] D. James and M. Philip, "a N Ovel a Nti P Hishing Framework Based on," vol. 3, no. December 2013, pp. 207–218, 2012.
- [9] B. L. To, L. A. T. Nguyen, H. K. Nguyen, and M. H. Nguyen, "A novel fuzzy approach for phishing detection," *2014 IEEE 5th Int. Conf. Commun. Electron. IEEE ICCE 2014*, pp. 530–535, 2014, doi: 10.1109/CCE.2014.6916759.
- [10] C. Wu, "2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA) A Phishing Detection System based on Machine Learning," *2019 Int. Conf. Intell. Comput. its Emerg. Appl.*, pp. 28–32, 2019.
- [11] S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A comparison of machine learning techniques for phishing detection," *ACM Int. Conf. Proceeding Ser.*, vol. 269, pp. 60–69, 2007, doi: 10.1145/1299015.1299021.
- [12] A. Subasi and E. Kremic, "Leveraging AI and machine learning for societal challenges, cas 2019 comparison of adaboost with multiboosting for phishing website detection," *Procedia Comput. Sci.*, vol. 168, no. 2019, pp. 272–278, 2020, doi: 10.1016/j.procs.2020.02.251.
- [13] H. S. Hota, A. K. Shrivastava, and R. Hota, "An Ensemble Model for Detecting Phishing Attack with Proposed Remove-Replace Feature Selection Technique," *Procedia Comput. Sci.*, vol. 132, pp. 900–

- 907, 2018, doi: 10.1016/j.procs.2018.05.103.
- [14] A. J. Obaid, K. K. Ibrahim, A. S. Abdulbaqi, and S. M. Nejrs, “An adaptive approach for internet phishing detection based on log data,” *Period. Eng. Nat. Sci.*, vol. 9, no. 4, pp. 622–631, 2021, doi: 10.21533/pen.v9i4.2398.
- [15] N. Sanglerdsinlapachai and A. Rungsawang, “Using domain top-page similarity feature in machine learning-based web phishing detection,” *3rd Int. Conf. Knowl. Discov. Data Mining, WKDD 2010*, pp. 187–190, 2010, doi: 10.1109/WKDD.2010.108.
- [16] M. Almseidin, A. M. Abu Zuraiq, M. Al-kasassbeh, and N. Alnidami, “Phishing detection based on machine learning and feature selection methods,” *Int. J. Interact. Mob. Technol.*, vol. 13, no. 12, pp. 71–183, 2019, doi: 10.3991/ijim.v13i12.11411.
- [17] N. Kumar, S. Sonowal, and Nishant, “Email Spam Detection Using Machine Learning Algorithms,” *Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2020*, no. September, pp. 108–113, 2020, doi: 10.1109/ICIRCA48905.2020.9183098.
- [18] S. Y. Yerima and M. K. Alzaylaee, “High Accuracy Phishing Detection Based on Convolutional Neural Networks,” *ICCAIS 2020 - 3rd Int. Conf. Comput. Appl. Inf. Secur.*, no. Iccais, pp. 19–21, 2020, doi: 10.1109/ICCAIS48893.2020.9096869.
- [19] M. Sánchez-Paniagua, E. Fidalgo, V. González-Castro, and E. Alegre, “Impact of current phishing strategies in machine learning models for phishing detection,” *Adv. Intell. Syst. Comput.*, vol. 1267 AISC, no. May, pp. 87–96, 2021, doi: 10.1007/978-3-030-57805-3_9.
- [20] A. Aljofey, Q. Jiang, Q. Qu, M. Huang, and J. P. Niyigena, “An effective phishing detection model based on character level convolutional neural network from URL,” *Electron.*, vol. 9, no. 9, pp. 1–24, 2020, doi: 10.3390/electronics9091514.
- [21] S. Nandhini and D. J. Marseline, “Performance Evaluation of Machine Learning Algorithms for Email Spam Detection,” *Int. Conf. Emerg. Trends Inf. Technol. Eng. ic-ETITE 2020*, pp. 1–4, 2020, doi: 10.1109/ic-ETITE47903.2020.312.
- [22] A. Hannousse and S. Yahiouche, “Towards benchmark datasets for machine learning based website phishing detection: An experimental study,” *Eng. Appl. Artif. Intell.*, vol. 104, pp. 1–21, 2021, doi: 10.1016/j.engappai.2021.104347.
- [23] R. Mahajan and I. Siddavatam, “Phishing Website Detection using Machine Learning Algorithms,” *Int. J. Comput. Appl.*, vol. 181, no. 23, pp. 45–47, 2018, doi: 10.5120/ijca2018918026.
- [24] X. Li, G. Geng, Z. Yan, Y. Chen, and X. Lee, “Phishing detection based on newly registered domains,” *Proc. - 2016 IEEE Int. Conf. Big Data, Big Data 2016*, pp. 3685–3692, 2016, doi: 10.1109/BigData.2016.7841036.
- [25] P. A. Barraclough, M. A. Hossain, M. A. Tahir, G. Sexton, and N. Aslam, “Intelligent phishing detection and protection scheme for online transactions,” *Expert Syst. Appl.*, vol. 40, no. 11, pp.

- 4697–4706, 2013, doi: 10.1016/j.eswa.2013.02.009.
- [26] H. Tupsamudre, A. K. Singh, and S. Lodha, “Everything is in the name – a URL based approach for phishing detection,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11527 LNCS, no. August, pp. 231–248, 2019, doi: 10.1007/978-3-030-20951-3_21.
- [27] “APWG | Phishing Activity Trends Reports,” *Apwg.org*, 2019. <https://apwg.org/trendsreports/>, last accessed on 29-12-2022 at 10:15 AM
- [28] Wikipedia Contributors, “Phishing,” *Wikipedia*, Feb. 01, 2019. <https://en.wikipedia.org/wiki/Phishing>, last accessed on 10-12-2022 at 06:15 PM
- [29] “What is Logistic Regression?” *TIBCO Software*. <https://www.tibco.com/reference-center/what-is-logistic-regression>, last accessed on 27-12-2022 at 10:56 AM.
- [30] B. Alam, “Implementing Decision Tree Using Python,” *Hands-On-Cloud*, Jan. 12, 2022. <https://hands-on.cloud/decision-tree-using-python/>, last accessed on 27-12-2022 at 11:02 AM
- [31] “Fig. 3 A general architecture of XGBoost,” *ResearchGate*. https://www.researchgate.net/figure/A-general-architecture-of-XGBoost_fig3_335483097, last accessed on 27-12-2022 at 11:07 AM
- [32] “What is Random Forest? | IBM,” *www.ibm.com*. <https://www.ibm.com/topics/random-forest>, last accessed on 27-12-2022 at 11:10 AM
- [33] Sunil, “Understanding Support Vector Machine algorithm from examples (along with code),” *Analytics Vidhya*, Mar. 11, 2019. <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>, last accessed on 27-12-2022 at 11:15 AM

AN_EFFECTIVE_APPROACH_FOR_PHISHING_DETECTION_USIN...

ORIGINALITY REPORT

22%

SIMILARITY INDEX

19%

INTERNET SOURCES

7%

PUBLICATIONS

13%

STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	9%
2	Submitted to Daffodil International University Student Paper	3%
3	www.mdpi.com Internet Source	1%
4	Submitted to Curtin University of Technology Student Paper	1%
5	www.hindawi.com Internet Source	<1%
6	www.researchgate.net Internet Source	<1%
7	madeyski.e-informatyka.pl Internet Source	<1%
8	Che-Yu Wu, Cheng-Chung Kuo, Chu-Sing Yang. "A Phishing Detection System based on Machine Learning", 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA), 2019 Publication	<1%