

**SPAM EMAIL DETECTION USING MACHINE LEARNING**

**BY**

**Afea Ufat Niha  
ID: 191-15-12730**

**AND**

**Mahedi Hasan Shovo  
ID: 191-15-13012**

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

Supervised By

**Mr. Fahad Faisal**

Assistant Professor  
Department of CSE  
Daffodil International University

Co-Supervised By

**Mr. Md. Sazzadur Ahamed**

Assistant Professor  
Department of CSE  
Daffodil International University



**DAFFODIL INTERNATIONAL UNIVERSITY**

**DHAKA, BANGLADESH**

**JANUARY 2023**

## **APPROVAL**

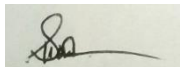
This Project titled “**SPAM EMAIL DETECTION USING MACHINE LEARNING**”, submitted by Afea Ulfat Niha 191-15-12730 and Mahedi Hasan Shovo 191-15-13012 to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment 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 26 January, 2023

## **BOARD OF EXAMINERS**

**Chairman**

---

**Dr. Touhid Bhuiyan**  
**Professor and Head**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**Internal Examiner**

---

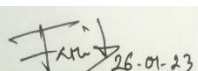
**Subhenur Latif**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**Internal Examiner**

---

**Mohammad Monirul Islam**  
**Assistant Professor**  
Department of Computer Science and Engineering  
Faculty of Science & Information Technology  
Daffodil International University



**External Examiner**

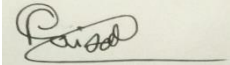
---

**Dr. Dewan Md Farid**  
**Professor**  
Department of Computer Science and Engineering  
United International University

## DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Mr. Fahad Faisal, 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 award of any degree or diploma.

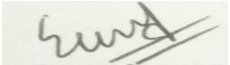
### Supervised by:



---

**Mr. Fahad Faisal**  
Assistant Professor  
Department of CSE  
Daffodil International University

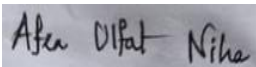
### Co-Supervised by:



---

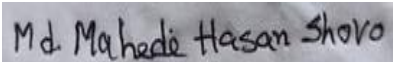
**Mr. Md. Sazzadur Ahamed**  
Assistant Professor  
Department of CSE  
Daffodil International University

### Submitted by:



---

**Afea Ulfat Niha**  
ID: -191-15-12730  
Department of CSE  
Daffodil International University



---

**Mahedi Hasan Shovo**  
ID: -191-15-13012  
Department of CSE  
Daffodil International University

## ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty ALLAH 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 **Mr. Fahad Faisal, Assistant Professor**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Data Mining & 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 drafts and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, Head, Department of CSE, for his kind help to finish our project and also to other faculty member 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**

Machine learning is a limb of artificial intelligence (AI) and information technology. Supervised machine learning is a subcategory of artificial intelligence and also the machine learning which is branch of (AI). The supervised learning always utilizes on labeled data that can train algorithm to accurately predict the results and classify data. Now a days day by day technology is more positive and negative also. In that the negativity one is spam mail which is under phishing and this can be recover with machine learning algorithms because it provides a perfect test result for detect spam mail. The goal of this study case that here describes the way that detect spam mail through the algorithms and applying them visually. Here we preferred Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) algorithms. Then after testing Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes these all algorithm gives this accuracy frequently 98 %, 96% and 0%. The rest of one algorithm is discuss with figures. After testing data Support Vector Algorithm gives highest accuracy which is 98%

## TABLE OF CONTENTS

<b><u>CONTENTS</u></b>	<b><u>PAGE</u></b>
Board of examiners.....	i
Declaration.....	ii
Acknowledgements .....	iii
Abstract.....	iv
<b>CHAPTER 1:</b>	
<b>INTRODUCTION</b>	<b>1-4</b>
1.1 Introduction	1
1.2 Motivation	3
1.3 Objectives	3
1.4 Research Questions	3
1.5 Expected Outcomes	4
<b>CHAPTER 2: BACKGROUND</b>	<b>5-9</b>
2.1 Introduction	5
2.2 Related Works	6
2.3 Comparative Studies	8
2.4 Research Scope	9
2.4 Challenges	9
<b>CHAPTER 3: RESEARCH METHODOLOGY</b>	<b>10-16</b>
3.1 Introduction	10
3.2 Research Subject and Instrumentation	10
3.3 Data Collection	10

3.4 Statistical Analysis	11
3.5 Data Pre-Processing	11
3.6 Proposed Model Workflow	12
3.7 Data Insertion	13
3.8 Split Data	13
3.9 Machine Learning Model	13
3.9.1 Support Vector Machine (SVM)	15
3.9.2 Logistic Regression (LR)	14
3.9.3 Naïve Bayes (NB)	14
3.9.4 Random Forest (RF)	15
3.10 Implementation Requirements	16
<b>CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION</b>	<b>17-21</b>
4.1 Introduction	17
4.2 Experimental Results	17
4.3 Descriptive Analysis	18
4.4 Comparison of accuracy	21
<b>CHAPTER 5: CONCLUSION AND FUTURE SCOPE</b>	<b>22-23</b>
5.1 Conclusion	22
5.2 Scope for Further Studies	23
<b>REFERENCES</b>	<b>24</b>
<b>PLAGIARISM REPORT</b>	<b>25</b>

## LIST OF FIGURES

<b>FIGURES</b>	<b>PAGE NO</b>
Figure 1.1: Common Email Spam File Type	1
Figure 1.1.2: Simple Spam Mail Detection Method	2
Figure 2.1: Common Email Spam	5
Figure 3.6.1: Proposed Model Workflow	12
Figure 3.7.1: Data Insertion	13
Figure 3.9.1.1: Model of SVM	14
Figure 3.9.2.2: Model of LR	15
Figure 3.9.3.3 Model of NB	15
Figure 3.9.4.4 Model of RF	16
Figure 4.3.1: Test Result of SVM	18
Figure 4.3.2: SVM Data Separate	19
Figure 4.3.3: Test Result of LR	19
Figure 4.3.4: Detect email	20
Figure 4.3.5: Detect Error Spam Email	20
Figure 4.3.6: Predict from Raw Data	21



## LIST OF TABLES

<b>TABLES</b>	<b>PAGE NO</b>
Table 2.1: Comparison between previous relative works	9
Table 3.1: Dataset Graph	11
Table 3.2: Information of Mails	11
Table 4.1: Report of Proposed Model	17
Table 4.4.1: Test accuracy of comparative analysis	21

# CHAPTER 1

## INTRODUCTION

### 1.1 Introduction

Spam mail is a branch of phishing. Spam mail problem is the smallest field of phishing. Phishing is what that steal data or information by spreading fake links or URL via mails, text messages. Spam mail is known as mail spam, junk mail, spam and so on. All of the email users are definitely face this spam message problem. There is no way to detect through normal visualization. Detect spam mail messages is positively dangerous to manage. The worldwide of the report basis on January 2021 there are 122.33 (billions) spam emails along the number of 144.76 (billion). As the worldwide report in 2022 there are 333.22 billion mails are sent per day and between them approximately 88.9 billion messages are spam. This spam emails are harmful for the device users and also for the devices.

Machine learning is a field that can help to detect the spam messages by using algorithm with 99.9 % accuracy. Machine learning is a part of artificial intelligence where supervised machine learning contain labeled dataset. Actually, supervise machine learning algorithm can assist us to free from spam messages. There comes to many innominate messages through spam messages which target was to attract the users to catch them in net and takes all secret information of the users. So, there is nothing but also apply machine learning algorithms to detect the worldwide problem – spam mail problems. There have so perfect algorithms to prevent this problem of this [1].

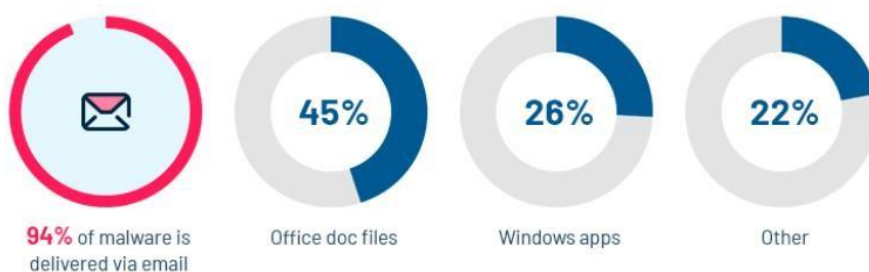


Figure 1.1: Common Email Spam File Type

According to this figure 1.1 here shown that 94% spam is come from delivered via mail which is the big portion agent of delivered spam emails and the rest of only 6% came from another root.

In this study, we proposed a scheme which based on machine learning Support Vector Machine (SVM), Logistic Regression (LR) model to detect spam email with a good accuracy. Beside this Naïve Bayes, Random Forest (RN) also apply here different way. Our study is capable of reducing spam email to provide good accuracy with effective workflow diagram. Due to detect spam messages we able to detect it approach with labeled data with operative algorithm models [2].

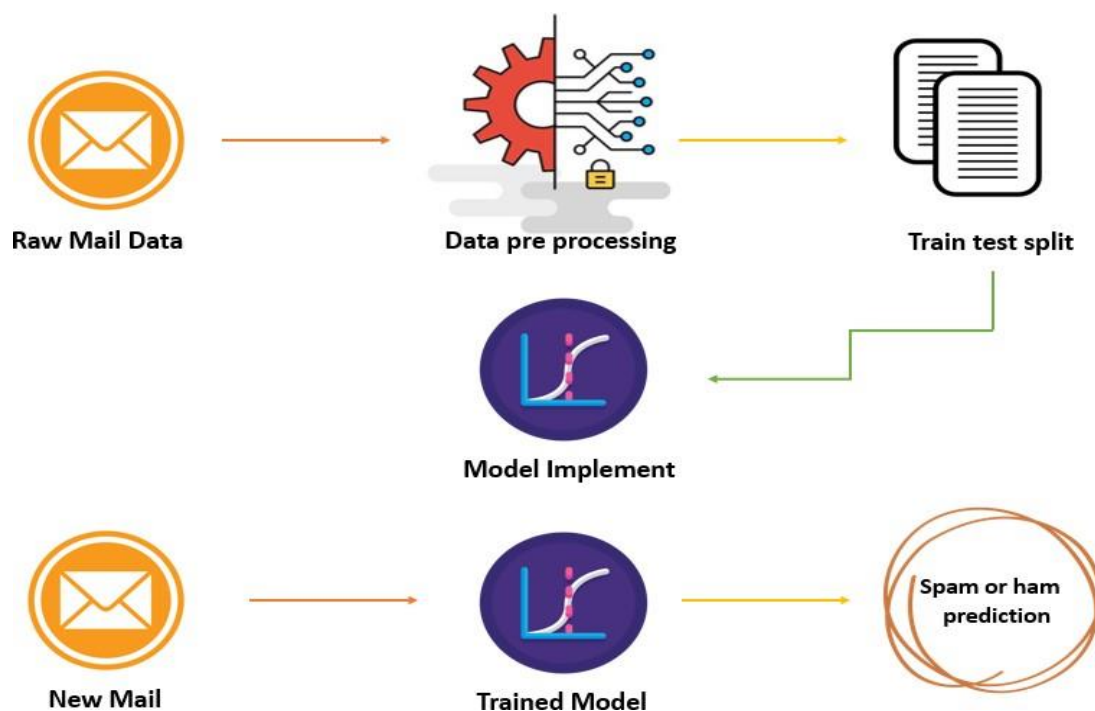


Figure 1.1.2: Simple Spam Mail Detection Method

According to this paper, here discuss how to prevent spam email problem use with different algorithm. As a result, we easily got a conclusion of detecting spam mail messages by sorting collecting dataset via implement machine learning classification algorithms because it can be predicted emails which is spam or not. Of course, data pre-processing is very weighty to workout good with algorithms.

## 1.2 Motivation

Machine learning is a system that work on datasets, some are labeled or some unlabeled data. In this paper applying Supervised Machine Learning. Supervised Machine Learning is subdivision of machine learning. By using supervised learning algorithm can be train the datasets as well they predict from dataset also. It turns datasets into factual. It works on labeled data by choosing model. Labeled data means (text files, images and so on). Then fit the model. After evaluate the model find out the perfect prediction with accuracy. In point of fact, going to this method we invent a desire outcome and the best approximates between training and testing data. For this work we conduct extensive length of Supervised Machine Learning labeled data implement with preferred Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes, Random Forest algorithms. By using this we detect spam messages from raw mails with 98% best accuracy [3] [4] [5] [6].

One motivation for using supervised learning for spam email detection is that it can be more effective than other methods, such as rule-based approaches or simple heuristics. Supervised learning algorithms can learn complex patterns in the data and make more accurate predictions as a result. Additionally, supervised learning models can be fine-tuned and improved over time as more data becomes available.

## 1.3 Objective

For problem solving machine learning one of the best methods. Supervised Machine Learning is a class of machine learning which helps to filtering mails in various categories like spam mail message. Then train the dataset (dataset must be labeled data) through the effective models and able to determine to found this message spam or ham message. It is the mainly goal of supervised machine learning. This detection is possible to applying machine learning.

## 1.4 Research Questions

- How can we collect dataset?
- Which machine learning model used in this experiment?
- What is the procedure of data pre-processing?
- Which method perform well in this case?
- What is the entrance to detecting spam email?
- Which algorithm is better for detecting spam email?
- What is the accuracy of best proposed model?

## **1.5 Expected Outcomes**

In this paper, the expected outcome of spam mail detection by using supervised learning is that the model will be able to accurately classify emails as either spam or non-spam based on its training data. The model will be able to identify patterns and features that are commonly associated with spam emails, and use this information to make predictions about the likelihood that a given email is spam. The accuracy of the model will depend on the quality of the training data and the effectiveness of the chosen supervised learning algorithm. We apply machine learning approaches to detect spam email. By using models, we find perfect result with high accuracy via predict from datasets to detect spam email. The outcomes of these algorithms have then been compared in terms of accuracy, error rate, precision, and recall for a perfect report.

## CHAPTER 2 BACKGROUND

### 2.1 Introduction

Spam mail detection is applying with machine learning technique. For spam messages we construct data from Kaggle raw mail data for our newly experiment. The main motive of our paper is to detect spam messages to embed with algorithm models with prediction high accuracy for better perform. We execute here some new formula for best model work with outcomes. Then we research some earlier work to relate with our proposed plan and so on.

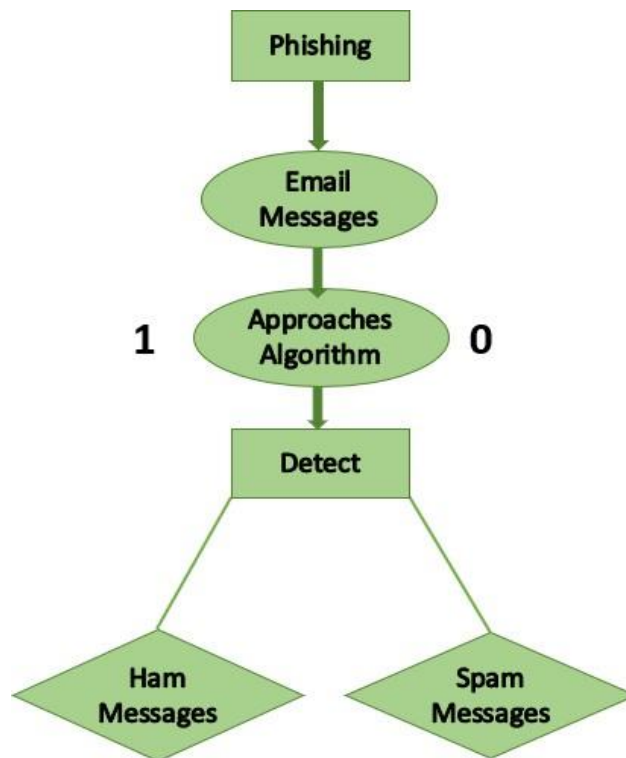


Figure 2.1: Common Email Spam

## 2.2 Related Work

Author Junaid Rashid, proposed different spam phishing by applying algorithm of machine learning. Support vector machine classifier, performance is best of 95.66% of spam phishing the proposed method exhibits results when range of standard spam phishing datasets. The proposal of paper there used SVM 95.66% accuracy and gives very low false-positive value. The new proposed method can detect spam phishing sites by applying machine learning method [7].

Noor Faisal Abedin, he arranged to collect data of a user without consent through emails. Via spam emails an attacker collect essential information of the user. After studies have discuss possible approaches to detect spam phishing. This paper work with three machine learning algorithms to detect spam phishing status. The random forest (RN) classifier performed well of 97%, a recall 99% and F1 Score is 97%. Proposed model work fast and outcome shows perfect as well. After the very good performance in future, they want to again change features and improve the accuracy more to pick more perfect dataset than the critical one [1].

Author Che-Yu Wu, as spam phishing pages to detect the spam phishing. In this article, they propose a detection system which combining the URL. In this they implement Support-vector machine model The system is provide high accuracy and low false positive rate detection results for spam phishing pages [8].

Fatima Salahdine, proposed a spam phishing attack detection technique using machine learning. Here used 3 models that are trained and tested on the dataset. For each classifier, SVM is provide high accuracy. For LR, also give high accuracy is given by corresponding to 0.4. For ANN, high accuracy is achieved with two hidden layers, In the end, the proposed model shows fast and accurate spam phishing attack detection [3].

Author John Arthur, presents a review of spam phishing detection. This paper discussed and applying six algorithm used classification methods of Machine Learning. The Machine Learning methods chosen Naïve Bayes (NB), Super Vector Model (SVM), Decision Tree (DT), Random Forest (RF), k-means clustering, and ANN. Based on the discussion of the ML methods, it is quite hard for them to determine the best one for advantages. In future they overcome this problem [6].

Merlin.V. Kunju, the main objective of this paper is to survey of spam phishing and its detection. This study here uses also machine learning techniques such as KNN Algorithm, Naïve Bayes (NB), Decision Tree (DT), Support Vector Machines (SVM), Neural Network (NN) and Random Forest (RF) algorithm for detect spam phishing information. In this research also given Comparison table which been prepared to tally the advantages, drawbacks, methodologies of the various approaches. By using each technique, they cannot carry a good decision so in future to developed this experiment again [2].

Author Saeed Abu-Nimeh, in paper present study compares Logistic Regression (LR), Classification and Regression Trees (CART), Bayesian Additive Regression Trees (BART), Support Vector Machines (SVM), Random Forests (RF), and Neural Networks (NNet) for detecting phishing emails. The results of future work are to explore more for perfect accuracy with this experiment [5].

Vahid Shahrivari, proposed one of the most successful methods for detecting these Spam phishing via Machine Learning. In this paper, they likened the results of multiple machine learning methods Logistic Regression (LR), Decision Tree (DT), Support Vector Machine (SVM), Ada Boost, Random Forest (RF), Neural Networks (NN), KNN, Gradient Boosting, and XG Boost. According to the result very good performance in Random Forest, XG Boost both [9].

Al Maha Abu Zuraiq, review different spam phishing detection approaches which is Content-Based, Heuristic-Based, and Fuzzy rule-based approaches. But using all these approaches that can't conclude a perfect approach to be used in detecting spam phishing by using machine learning [4].

Author Sameena Naaz, review that the outcome has been compared with previous work with same dataset. The outcomes of these algorithms have been compared of accuracy, error rate, precision, and recall. In this paper work different algorithms have been compared on spam phishing dataset. Then the Random Forest works better and give very good accuracy, error rate [10].

Mohammed Almseidin, chosen efficiency of the models. Random forest (RF) gives the best result of the experiment achievement and the accuracy was 98.11% [11].



Author Neda Abdelhamid, review the purpose of the advantages and disadvantages of ML predictive models. As a result, that Covering approach models are more appropriate for spam phishing solutions but in the near future they want to apply Super Vector Machine (SVM) for detecting this experiment [12]

### 2.3 Comparative Studies

Now a days spam phishing email problem increasing that steal all personal information that's why experimental process is help out to detect this problem. After study all relative works then some short comparison note is making through the below table 2.1 that what work was done and what work or experiment that we'll provide.

Table 2.1: Comparison between previous relative works

References	Title	Algorithms	Best Accuracy
[7]	Phishing Detection Using Machine Learning Technique	SVM	95.66%
[1]	Spam Detection using Machine Learning Classification Techniques	RF	97%
[8]	Phishing Detection System based on Machine Learning	SVM	High
[3]	Spam Phishing Attacks Detection MachineLearning-Based Approach	LR, ANN, SVM	High
[6]	Review of the machine learning methods in the classification of phishing attack	SVM, DT, LR ANN, NB	Not defined (Future Work)
[2]	Evaluation of Phishing Techniques Based on Machine Learning	SVM, DT, LR ANN, NN, RF, NB	Not defined (Future Work)
[5]	A Comparison of Machine Learning Techniques for Phishing Detection	SVM, LR, CART, BART, RF	Not defined (Future Work)
[9]	Spam Phishing Detection Using Machine Learning Techniques	XG Boost , RF	High
[4]	Review: Phishing Detection Approaches	Content-Based, Heuristic-Based, and Fuzzy rule	Not defined

			(Future Work)
[10]	Detection of Phishing in Internet of Things Using Machine Learning Approach	RF	High
[11]	Phishing Detection Based on Machine Learning and Feature Selection Methods	RF	98.11%
[12]	A Recent Intelligent Machine Learning Comparison based on Models Content and Features	SVM	Not defined (Future Work)

## 2.4 Research Scope

In our research, to using Machine learning model by applying methods which gives very good accuracy and also the error rate. By using different type of algorithms and classifiers we train our model for better outcomes as well. Our purpose is to give better train our model for better work to detect the spam email problem.

## 2.5 Challenges

- Collecting data.
- Import data
- Data preprocessing and labeled data
- Data collection for research study.
- Getting accuracy above 90%.
- Making a decision based on the results of tests.

# CHAPTER 3

## RESEARCH METHODOLOGY

### 3.1 Introduction

It is in this; we represent our experiment system or methods that how collect dataset then from dataset how we'll done all steps of our experiment and how we implement all the model for the better accuracy and error rate. In this chapter we also proposed model workflow and all data work as well. So, for better and easy to get all information that we present all of in this chapter.

### 3.2 Research Subject and Instrumentation

This research here detects spam message with highest level of good performance as far as possible. We use here Supervised Machine Learning with labeled data to run our model. In supervised learning, some classification methods are used to run our model better. Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Naïve Bayes method are used in this research. Among all them Support Vector Machine (SVM) and Logistic Regression (LR) provide above 90% accuracy. At the end for selecting quality full dataset for a best outcome.

### 3.3 Data Collection

We collecting dataset for train actually we collect raw data which was reporting on TechVidvan website which is study based. Then we find out dataset from Kaggle on email data which contain two type email one is spam and another is ham. Ham means good email. Then labeled data into x and y. There total 5572 mails among them 747 spam messages and 4825 ham messages. After apply our model then we find out some duplicate data that was removed. The collecting data graph Table 3.1 is given below.

Table 3.1: Dataset Graph

		Category			
		Count	ham	spam	Total
Category	ham	4825.0	0.0	<b>4825.0</b>	
	spam	0.0	747.0	<b>747.0</b>	
	Total	<b>4825.0</b>	<b>747.0</b>	<b>5572.0</b>	

### 3.4 Statistical Analysis

We have collected our data from Kaggle. It has 5572 records and 2 columns (ham and spam).

Category: Ham, Spam

Messages: 5572

Table 3.2: Information of Mail

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

### 3.5 Data Pre-processing

At first, we collect data then label encode data and split into x and y. The dataset has 5572 amounts of data. Then splitting the data into training and testing data. After train our model at the end we find above 90% better outcome.

### 3.6 Proposed Model Workflow

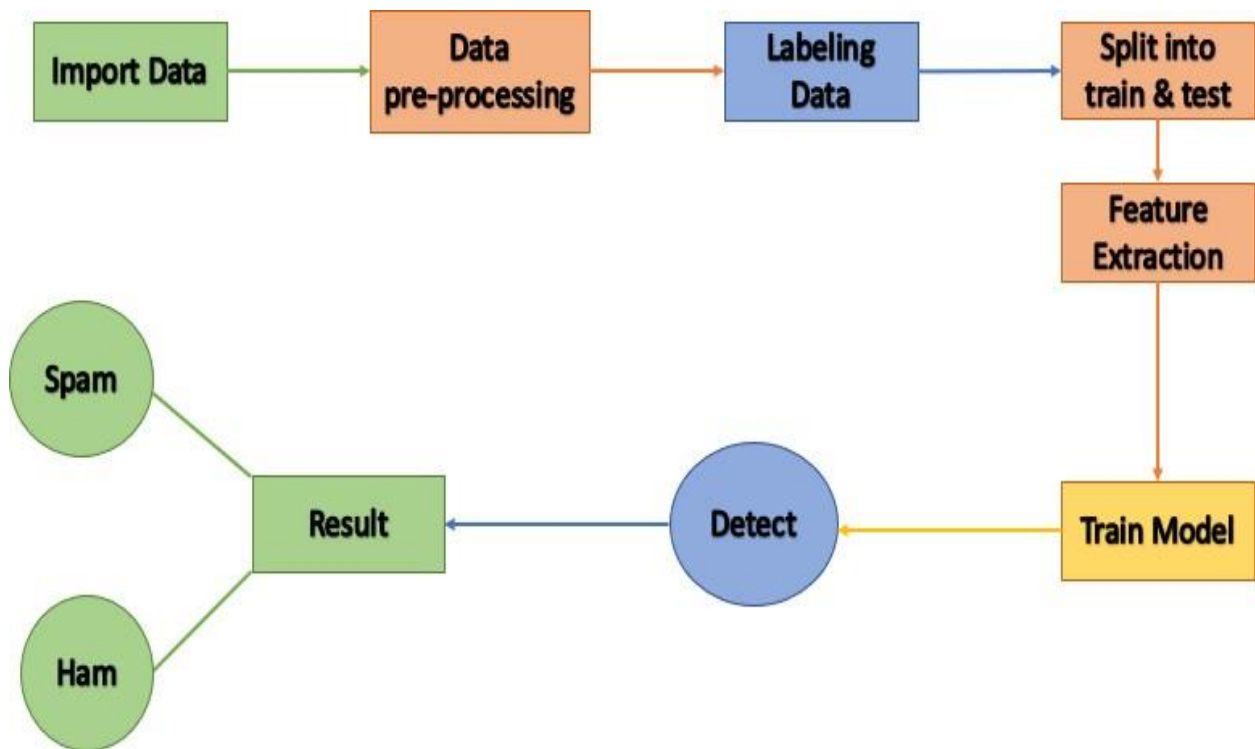


Figure 3.6.1: Proposed Model Workflow

In Figure 3.6.1 train models are - Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF). Here at first import dataset then pre-process data. After that labeling the data and split into train and test. Do some feature extraction for train our model with algorithms. By applying algorithms, it can be detected as spam email and other will ham email.

### 3.7 Data Insertion

After select quality data then input the dataset and connect with the software where we implement. Then after connecting we see that our data insertion is ok or not.

```
# loading the data from csv file to a pandas Dataframe
raw_mail_data =pd.read_csv('/content/mail_data.csv')
```

```
print(raw_mail_data)
```

	Category	Message
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
...	...	...
5567	spam	This is the 2nd time we have tried 2 contact u...
5568	ham	Will ü b going to esplanade fr home?
5569	ham	Pity, * was in mood for that. So...any other s...
5570	ham	The guy did some bitching but I acted like i'd...
5571	ham	Rofl. Its true to its name

Figure 3.7.1: Data Insertion

### 3.8 Split Data

Whenever data insertion is complete then we split the dataset into two parts one is training dataset and another is testing dataset which contain x and y. Training dataset is the raw data which we find from our main dataset and test dataset is the minor dataset that conduct to test our model.

### 3.9 Machine Learning Model

Machine learning algorithms (Supervised Machine Learning) used in text classification, detection, image filtering and so on. Using this machine learning algorithms technique, we detect spam email from the ham messages. In this experiment we preferred Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) algorithms models to detect spam mails from raw or ham mails.

### 3.9.1 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a classifier of Machine Learning. It is supervised learning classifier and also contain labeled data. SVMs are mainly utilized handwriting recognition, image detection, face detection, email classification, gene classification, and so on. There are two types of (SVM). SVM can handle both linear and non-linear data as well. SVM have effective at high altitudes.

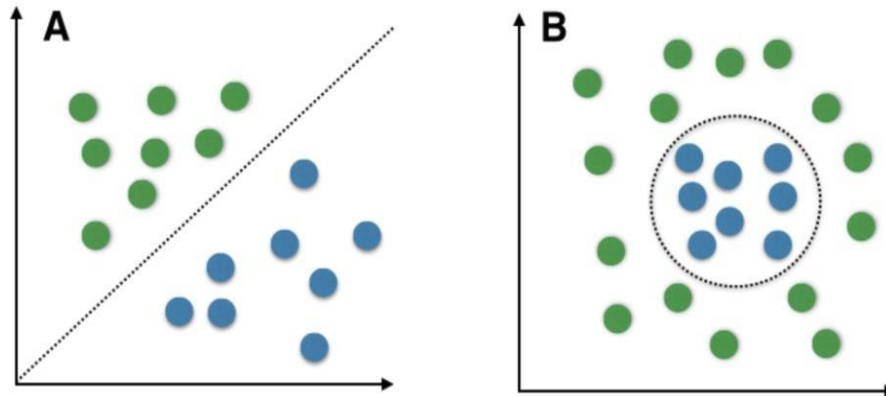


Figure 3.9.1.1: Model of SVM

### 3.9.2 Logistic Regression (LR)

Logistic Regression (LR) is Supervised Machine Learning which contain labeled data. (LR) predict the probability, or calculate the outcome yes, no; truth, false so on. It can predict whether or not an email is a spam, detect image, text detect and many other.

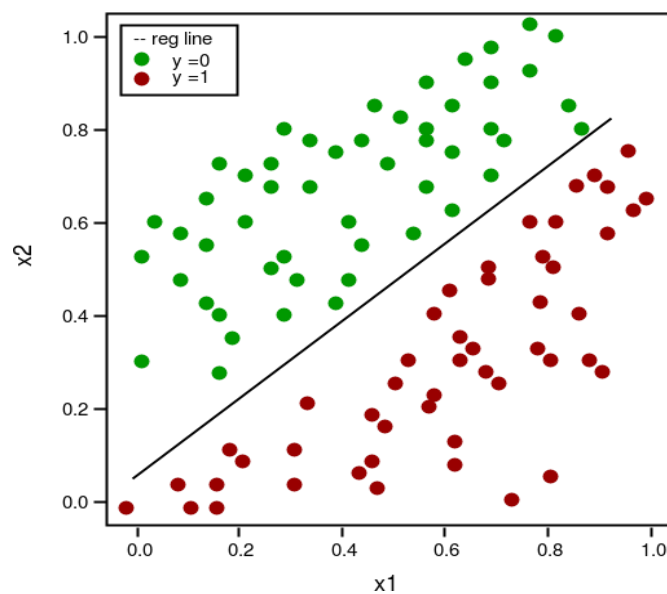


Figure 3.9.2.2: Model of LR

### 3.9.3 Naïve Bayes (NB)

Naïve Bayes (NB) is a machine learning model utilize on large amounts of labeled data. It is simple algorithm use in Bayes rule. It calculates the probability of each class and then pick one with the highest probability.

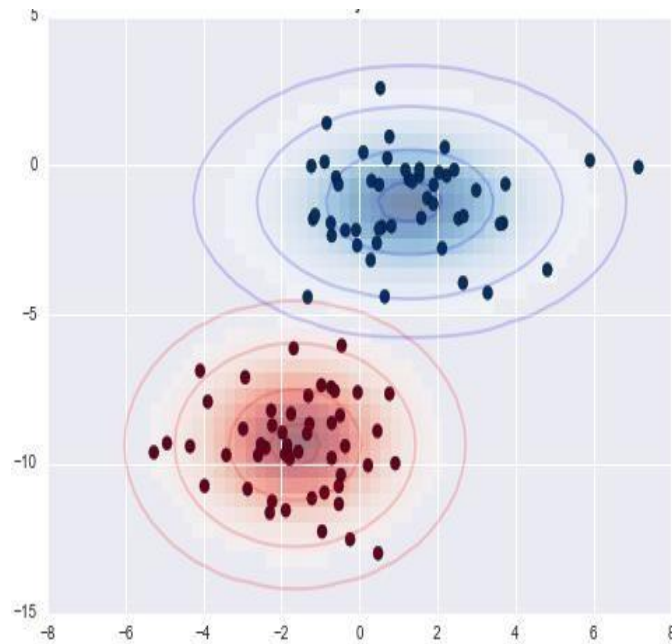


Figure 3.9.3.3 Model of NB

### 3.9.4 Random Forest (RF)

Random Forest (RF) is supervised machine learning algorithm with labeled data to solving all regression problems. It is the match with decision trees.



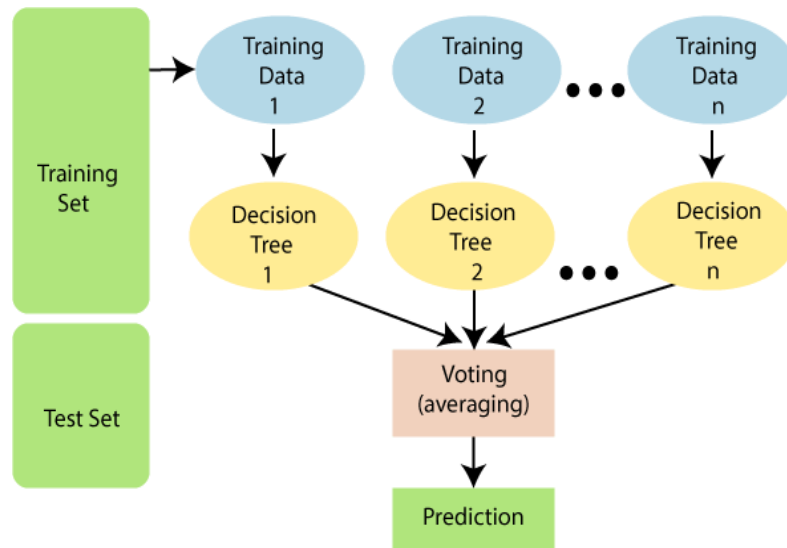


Figure 3.9.4.4 Model of RF

### 3.10 Implementation Requirements

We use some hardware, software and some developing tools for our experiment.

Hardware, Software requirement –

- Operating System (Windows 7,8,10,11)
- Web Browser (chrome or firefox)
- Ram 4 GB
- Hard drive 500 GB
- Processor intel CORE i3, i5

Developing tools –

- Python
- Colab
- Orange

## CHAPTER 4

### EXPERIMENTAL RESULTS AND DISCUSSION

#### 4.1 Introduction

In Chapter 4 we discuss here the experimental result of our proposed data, descriptive analysis of the data and the graphical outcome of our models.

#### 4.2 Experimental Results

In this, we proposed here the expected outcomes of our model which use this experiment. We use Machine Learning technique use supervised learning with labeled data. Here applying four model to detect spam mails from raw mails. Here we preferred Logistic Regression (LR), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) algorithms. Among these Logistic Regression (LR), Support Vector Machine (SVM) perform best than others. The accuracy rate of Support Vector Machine (SVM), Logistic Regression (LR) gradually 98.11% and 96%. Support Vector Machine (SVM) model make best accuracy which is above 95% then gradually Logistic Regression (LR) is 96% accuracy which also above 95% but not best than (SVM) as well. Table 4.1 shows the classification report of four classifiers which proposed on this model.

Table 4.1: Report of Proposed Model

<b>Algorithms</b>	<b>Accuracy</b>	<b>Error Rate</b>
SVM	98.11%	02.00
LR	96%	04.00
NB	0%	0.01
RF	0%	0.01

### 4.3 Descriptive Analysis

In this paper work, we found the best result is given by Support Vector Machine (SVM) 98% accuracy and secondly Logistic Regression (LR) 96% accuracy and best perform than other classifiers. All of the classifiers result and outcome are given below.

#### Support Vector Machine (SVM)

```
# Printing the results
print("Accuracy for SVM is:",accuracy)
print("Confusion Matrix")
print(confusion_mat)
```

Accuracy for SVM is: 97.96296296296296

Confusion Matrix

```
[[53  0  0  0  0  0  0  0  0  0]
 [ 0 50  0  0  0  0  0  0  1  0]
 [ 0  0 54  0  0  0  0  0  0  0]
 [ 0  0  0 54  0  1  0  0  0  0]
 [ 0  1  0  0 53  0  0  0  0  0]
 [ 0  0  0  0  0 57  0  0  0  1]
 [ 0  0  0  0  0  0 51  0  1  0]
 [ 0  0  0  0  0  0  0 54  1  0]
 [ 0  4  0  0  0  0  0  0 51  0]
 [ 0  0  0  0  0  0  0  0  1 52]]
```

Figure 4.3.1: Test Result of SVM

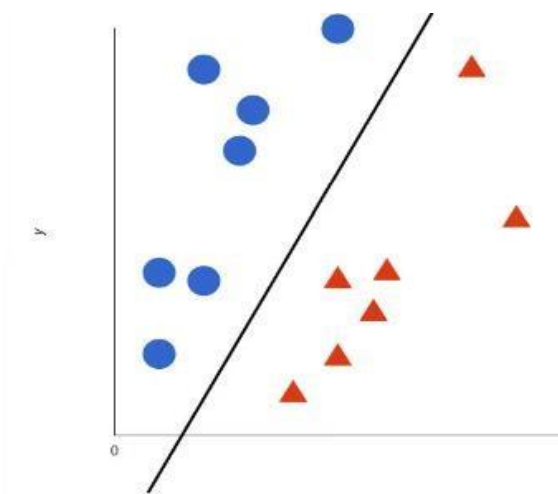


Figure 4.3.2: SVM Data Separate

## Logistic Regression (LR)

```
[ ] print('Accuracy on test data : ', accuracy_on_test_data)
```

```
Accuracy on test data : 96.9659192825112107
```

Figure 4.3.3: Test Result of LR

```

[ ] input_mail = ["I've been searching for the right words to thank you for this breather. I promise i wont take

# convert text to feature vectors
input_data_features = feature_extraction.transform(input_mail)

# making prediction

prediction = model.predict(input_data_features)
print(prediction)

if (prediction[0]==1):
    print('Ham mail')

else:
    print('Spam mail')

[1]
Ham mail

```

Figure 4.3.4: Detect email

## Naïve Bayes (NB)

```

[ ] emails = ['I HAVE A DATE ON SUNDAY WITH WILL!!, Had your mobile 11 months or more? U R entitled to Update to the

[ ] cv_emails = cv.transform(emails)

[ ] model.predict(cv_emails)

array(['#ERROR!'], dtype='<U910')

```

Figure 4.3.5: Detect error spam email

## Random Forest (RF)

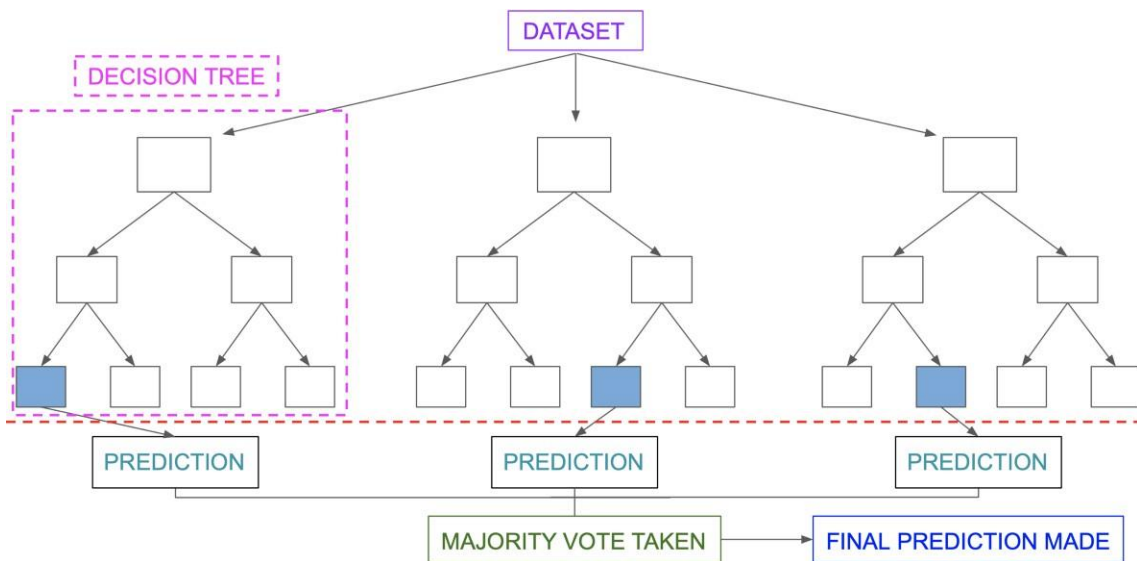
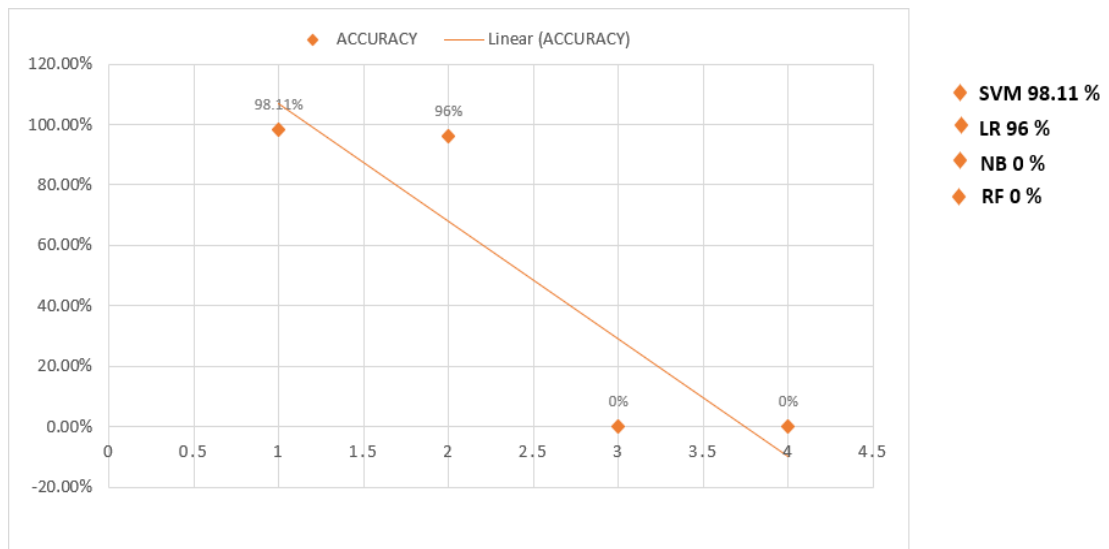


Figure 4.3.6: Predict from Raw Data

## 4.5 Comparison based on accuracy

Table 4.4.1: Test accuracy of comparative analysis



## CHAPTER 5

### CONCLUSION AND FUTURE SCOPE

#### 5.1 Conclusion

Spam mail problem has become the major issue that across over the internet security all over the world. The pit of spam mail problem is highly increasing gradually through many ways. One approach is to use a supervised learning algorithm, where the algorithm is trained on a labeled dataset of spam and ham emails. The algorithm can then be used to classify new emails as spam or ham. There are several factors that can influence the effectiveness of a supervised learning algorithm for spam mail detection. These include the quality and size of the training dataset, the choice of features and the type of algorithm used.

This research presents an overview of detect spam mail problem of our model. We propose four classifiers to detect spam mail and select the best one with best accuracy and perfect error rate. By applying SVM, our model performs best, beside this the another one LR, is also perform good as well. Each accuracy is above 95%. It's a good sign for our model.

#### 5.2 Scope for Further Studies

In future, to detect spam mail is probably going to concentrate on more classifier to implement this proposed model. Also try to give best perform from all selected classifier. Spam mail detection using machine learning is an active area of research, and there are many potential directions that future studies could take. Here are a few examples of potential future studies in this area:

**5.2.1 Developing new machine learning models:** There is always the possibility of developing new machine learning models that are more accurate or efficient at detecting spam emails. These models could be based on different types of data, such as text, images, or network metadata.

**5.2.2 Improving existing models:** Researchers could also focus on improving the performance of existing spam detection models, for example by developing more sophisticated feature engineering techniques or by incorporating additional types of data into the models.

**5.2.3 Evaluating the effectiveness of different techniques:** There may be a need to compare the effectiveness of different machine learning techniques for spam detection, in order to determine which methods are most effective in different contexts.

**5.2.4 Applying machine learning to new types of spam:** As spam evolves and becomes more sophisticated, it will be important to develop machine learning techniques that can effectively detect these new types of spam.

**5.2.5 Investigating the ethical implications of spam detection:** As machine learning techniques become more widespread, there may be a need to consider the ethical implications of using these techniques for spam detection, such as potential biases in the data or the impact on privacy.



## REFERENCES

- [1] R. B. T. S. M. S. M. A. R. S. H. Noor Faisal Abedin, "Phishing Attack Detection using Machine Learning Classification Techniques," *Proceedings of the Third International Conference on Intelligent Sustainable Systems*, pp. 1-6, 2020.
- [2] M. E. D. H. C. A. S. B. Merlin .V.Kunju, "Evaluation of Phishing Techniques Based on Machine Learning," *Proceedings of the International Conference on Intelligent Computing and Control Systems*, 2019.
- [3] Z. E. M. N. K. Fatima Salahdine, "Phishing Attacks Detection A Machine Learning-Based Approach".
- [4] M. A. AlMaha Abu Zuraiq, "Review: Phishing Detection Approaches," pp. 1-6.
- [5] D. N. X. W. S. N. Saeed Abu-Nimeh, "A Comparison of Machine Learning Techniques for phishing detection".
- [6] T. S. M. A. I. M. S. M. S. K. John Arthur Jupin, "Review of the machine learning methods in the classification of phishing attack," *Bulletin of Electrical Engineering and Informatics*, vol. 8, pp. 1-11, 2019.
- [7] T. M. M. W. N. T. N. Junaid Rashid, "Phishing Detection Using Machine Learning," *First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)*, pp. 1-4, 2020.
- [8] C.-C. K. C.-S. Y. Che-Yu Wu, "A Phishing Detection System based on Machine Learning," *International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, pp. 1-5, 2019.
- [9] V. S. M. I. Mohammad Mahdi Darabi, "Phishing Detection Using Machine Learning Techniques," vol. 1, pp. 1-9, 2020.
- [10] J. H. Sameena Naaz, "Detection of Phishing in Internet of Things Using Machine Learning Approach," *International Journal of Digital Crime and Forensics*, vol. 13, no. 2, pp. 1-15, 2021.
- [11] A. A. Z. M. A.-k. Mohammed Almseidin, "Phishing Detection Based on Machine Learning and Feature Selection Methods," pp. 1-13.
- [12] F. T. H. A.-j. Neda Abdelhamid, "Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features".

## Turnitin Originality Report

Processed on: 05-Jan-2023 23:04 +06  
 ID: 1988900768  
 Word Count: 4388  
 Submitted: 1

Phishing By Afea Ulfat Niha

Similarity Index

25%

Similarity by Source

Internet Sources: 21%  
 Publications: 12%  
 Student Papers: 12%

3% match (student papers from 13-Apr-2018)

[Submitted to Daffodil International University on 2018-04-13](#)

2% match (Internet from 06-Aug-2022)

<https://cra-vp.org/node/488/biblio/keyword/837?order=asc&page=24&sort=title>

2% match (Internet from 11-Sep-2022)

<https://academic-accelerator.com/Manuscript-Generator/Decision-Tree/Random-Forest-Classifer>

1% match (Internet from 20-Nov-2022)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/3495/013319%20%2B18%25%29.pdf?isAllowed=y&sequence=1>

1% match (Internet from 20-Nov-2022)

[http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5248/162-15-7944%20%2B5\\_%29.pdf?isAllowed=y&sequence=1](http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5248/162-15-7944%20%2B5_%29.pdf?isAllowed=y&sequence=1)

1% match (Merlin. V. Kunju, Esther Dainel, Heron Celestie Anthony, Sonali Bhelwa. "Evaluation of Phishing Techniques Based on Machine Learning", 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019)

[Merlin. V. Kunju, Esther Dainel, Heron Celestie Anthony, Sonali Bhelwa. "Evaluation of Phishing Techniques Based on Machine Learning". 2019 International Conference on Intelligent Computing and Control Systems \(ICCS\), 2019](#)

1% match (Fatima Salahdine, Zakaria El Mrabet, Naima Kaabouch. "Phishing Attacks Detection A Machine Learning-Based Approach", 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2021)

[Fatima Salahdine, Zakaria El Mrabet, Naima Kaabouch. "Phishing Attacks Detection A Machine Learning-Based Approach". 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference \(UEMCON\), 2021](#)

1% match (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)

[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](#)

1% match (Internet from 19-Dec-2022)

<https://www.slideshare.net/journalBFEI/review-of-the-machine-learning-methods-in-the-classification-of-phishing-attack>

1% match (AlMaha Abu Zuraiq, Mouhammd Alkasasbeh. "Review: Phishing Detection Approaches", 2019 2nd International Conference on new Trends in Computing Sciences (ICTCS), 2019)

[AlMaha Abu Zuraiq, Mouhammd Alkasasbeh. "Review: Phishing Detection Approaches". 2019 2nd International Conference on new Trends in Computing Sciences \(ICTCS\), 2019](#)

1% match (Internet from 28-Sep-2022)

<https://www.igi-global.com/gateway/article/272830>

1% match (Internet from 21-Oct-2021)

<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0258361>

1% match (Internet from 10-Nov-2022)

<https://downloads.hindawi.com/journals/scn/2022/4794752.pdf>

1% match (Internet from 28-Dec-2022)

<https://www.coursehero.com/file/133096205/18BIT0210-VL2020210503846-PE003pdf/>

1% match (Internet from 24-Nov-2022)

[https://repository.mines.edu/bitstream/handle/11124/176523/1/u\\_mines\\_0052E\\_12225.pdf?isAllowed=y&sequence=1](https://repository.mines.edu/bitstream/handle/11124/176523/1/u_mines_0052E_12225.pdf?isAllowed=y&sequence=1)