MLSMS-SpamDetection: SMS Spam Detection using Machine Learning

Dr.Diaa Salama AbdElminaam[1], Eng.Tarek Talaat Mohammed[2]
Ahmed Abdelrhman Gaber [3], Ahmed Ayman Mahmoud [4], Ahmed Mohamed Adel [5],
Bassel AbdelRahim AbdelMohsen [6], Ahmed Mohamed Abd Elsadek[7]
Faculty of Computer Science
Misr International University, Cairo, Egypt
diaa.salama[1], tarek.talaat[2]
ahmed2103540[3], ahmed2107685[4], ahmed2100403[5], bassel2107433[6],ahmed2103297[7]{@miuegypt.edu.eg}

***Abstract*—SMS spam has become a significant concern, particularly due to the rise of role that mobile phones play in our life and the subsequent surge in spam messages. These messages pose threats to user privacy and increase the risk of potential financial scams. This paper aims to address these challenges by proposing and evaluating machine learning algorithms specifically designed for to detect sms spam. These algorithms include SVM (Support Vector Machine), Naive Bayes, Decision-tree, KNN, and Logistic Regression. We utilized three datasets to evaluate the efficiency of the algorithms and the title of these datasets are SMS Spam Detection Application(spam.csv), Filtering mobile phone spam(spamraw.csv), and Spam SMS Classification(TrainDataset.csv). We applied each algorithm to each dataset and calculated their respective accuracy. The results indicated that the SVM (Support Vector Machine) performed the best in all of the dataset, while the KNN had the highest precision in all of the datasets, with the SVM (Support Vector Machine) emerging as a robust baseline for SMS spam detection. The study reveals that machine learning algorithms can effectively detect SMS spam, highlighting the need for tailored approaches in this domain. Further research should explore strategies to enhance feature space to achieve a more effective SMS spam filtering system. This may involve using more advanced natural language processing techniques or employing deep learning methods. This paper has demonstrated the effectiveness of machine learning algorithms in detecting SMS spam. By applying these algorithms to various datasets and evaluating their accuracy, we have established that the SVM (Support Vector Machine) suitable baseline methods for SMS spam detection. Future research should focus on refining these methods and developing more advanced strategies to detect and filter SMS spam effectively.**

Keywords: SMS spam detection; Machine Learning; Classification; Naïve Bayes; SVM(Support Vector Machine); Linear Regression; K-Nearest Neighbor(KNN); Decision trees; Logistic Regression; Accuracy; Precision; Recall; F1 Score.

I. INTRODUCTION

SMS spam detection is crucial in modern communication with the widespread use of Short Message Service (SMS). Differentiating between real messages and spam is vital for user privacy. SMS spam, marked by unsolicited messages promoting products or scams, necessitates strong detection methods. The distinction spams content analysis, machine learning, and behavioral patterns. Recognizing these differences is vital for accurate identification and filtration, safeguarding users from security threats and disruption, emphasizing advanced detection technologies.

The prevalence of SMS spam has surged, driven by the diminishing costs associated with spamming. Messaging charges have plummeted to below US$ 0.001 in some markets and are even entirely free in certain regions, reflecting the widespread adoption of SMS. Cloudmark statistics reveal significant regional disparities in mobile phone spam. In 2010, North America experienced less than 1% of SMS text as spam, AS certain countries of Asia witnessed a much higher rate, with up to 30% of messages being spam [source: Cloudmark stats]. This data underscores the global challenge posed by SMS spam and emphasizes the pressing need for effective countermeasures.[1]

When using machine learning, we aim to develop systems capable of learning and predicting based on past experiences. SMS spam detection, an important aspect of AI, is our main objective. To achieve this, we utilize machine learning algorithms on datasets containing labeled SMS messages to teach our systems how to recognize patterns in text messages and make predictions accordingly. The supervised learning approach is widely used in this context. By training our machine learning algorithms on a dataset where SMS texts are labeled spam or ham, we can effectively build models that can accurately differentiate between spam and legitimate messages. We split this dataset using the train test split, with the training set serving as a foundation for our model's learning process, while the test set checks the model's predictive accuracy. By implementing various algorithms such as SVM, K Nearest Neighbors (KNN), Decision Trees, Naive Bayes, and Logistic regression, we are able to thoroughly analyze SMS datasets, leading to the creation of highly accurate models. These algorithms contribute to the machine learning approach's robustness in classifying messages and predicting outcomes.

This paper makes a unique contribution to the study of SMS spam detection by leveraging three distinct datasets: "SMS Spam Detection Application," "Filtering Mobile Phone Spam," and "Spam SMS Classification." The research stands out in its examination of five machine learning algorithms—SVM (Support Vector Machine), Naïve Bayes, Decision Trees, K Nearest Neighbors (KNN), and Logistic Regression—customized explicitly for SMS spam detection. Through a comparative assessment, our goal is to pinpoint the most effective algorithm for each dataset, presenting original insights that contribute to the advancement of spam filtering methodologies in various scenarios, ensuring the research's distinctiveness and avoiding plagiarism.

The paper is structured into several sections. "Related Work" section provides insights into earlier research efforts and their findings in the field. Moreover, The "Methodology" section outlines the research approach, cover the acquisition and preparation of the dataset . The "Results" section presents the outcomes of the implemented algorithms, confirming the highest accuracy achieved. The "Conclusion" section summarizes the key findings and potential conclusion of the study. Acknowledgment towards all the supporting figures of this paper is presented in the sixth section.

II. Related Work

The field of SMS Spam Detection is mot new in our generation, a lot of people investigated the field and came up with variuos results. We will review and analyze relevant literature to inform our research. All the papers examined will be appropriately cited in the references section to ensure academic integrity and give credit to the main authors.

Luo GuangJun et al.[2] found out that Logistic Regression (LR) which studies the relationship between a set of independent variables and the dependent binary variables, it is very accurate with its predictions, with less errors than all tested algorithms. This algorithm scored a 99% accuracy. The goal from this research is to classify the insecurity of mobile message communication due to spam, which is considered a big problem that face most of people . And they used a "SMS spam collection dataset", that contains 5572, splitted into 4900 ham and 672 spam messages.

Alzahrani et al.[3] trained many algorithms to get best accuracy in SMS Spam detection. The aim from this program is to filterize the spam and ham sms messages because it considers one of the major problems that cell phone users face everyday, which recieve unwanted and mostly annoying SMS messages from their advertisers or other source. The program used a dataset collected for SMS spam research, which contains 5,574 SMS messages in English, only 747 are spam which labeled as 1 and the rest are ham which labeled as 0. After training ,testing and comparing, it found that the neutral network algorithm had the best accuarcy, which was nearly 98% with only 2 errors out of 979 spam messages.

Ajay Rana et al. [4] estimated that SVM classifier, is the best model to detect SMS spam messages, which is trying to reach the maximum separating hyperplane between the different classes available in the target feature. They tried another algorithms like, Naive Bayes and Maximum Entropy (MAXENT), but the best result was SVM with accuracy at 97.4%, which surpassed the other algorithms. The data-set was recorded with total 5574 messages; 4827 ham messages and 747 spam messages.

Sridevi Gadde et al.[5] studies the problem of detecting SMS spam beacuse of the increase of spam SMS and the need for an effective way to detect SMS spam. The dataset used in this research paper was obtained form UCI machine learning repository consist of 5572 SMS, the SMS messages are labeled "spam" or "ham". This research paper used machine learning and deep learning for example KNN, SVM, Navie Bayes, Decision tree and LSTM (Long short term memory). The results were evaluated based on precision, recall and accuracy, The SVM had an accuracy of 97% but LSTM had an accuracy of 98%.

Mehul Gupta et al.[6] The essential goal is to solve security issues in terms of protection of privacy and accessibility. They brought two data-sets, the first data-set splitted into 4827 ham and 747 spam with total 5574 sms messages, and the second data-set divided into 1000 sms messages for each. After data collection, data preprocessing, training and testing, the results claimed from the evaluation of the classifiers showed that CNN (Convolutional Neural Network) classifier fulfilled the best accuracy of 99.19% and 98.25% and AR value of 0.9926 and 0.9994 for the two datasets respectively. Convolutional Neural Network (CNN) consists of multiple layers, it uses filters by applying it to input image then extract features, the pooling layer downsamples the image to decrease computation, and the fully connected layer makes the final prediction

Fatima Zohra et al.[7] aims to test the effectiveness of different machine learning models for detecting SMS spam. The data set in the paper is the tiago's sms spam data set which is labeled "spam" or "ham", the data set was per-processed. They used these machine-learning including multilayer perceptron (MLP), SVM,random forst and KNN. The results documented in this paper are as follow MLP achieves the highest accuracy of 98.42% and outscores other algorithms in terms of precision, recalland F-measure.

Yuliya Kontsewaya et al.[8] addresses the increase in SMS

spam and the effectiveness of the machine-learning in solving this problem. This paper uses a data set obtained form the UCI machine learning repository, it consists of 5572 SMS labeled as either spam or ham. The paper uses the following machine-learning models like SVM, Naive Bayes, KNN,Decision tree and random forest. The results are based of the following accuracy, ROC area, f-measure and recall. Logistic Regression and Naive Bayes were the ones with best accuracy up to 99%. The paper concludes by proposing the potential of of algorithms or filtering methods to enhance the intelligence of spam detection classifiers.

Tiago A. Almeida et al.[1] addresses difficulties in combating mobile phone spam, including nonavailability of spam-filtering software. This paper evaluates the effectiveness of several machine learning algorithms when used to detect SMS spam. They use a variety of algorithms which include Naive Bayes, SVM, Minimum description length (MDL), KNN and C4.5. The results were as follow that the SVM achieved the best score among the rest of the algorithms.

Suparana DasGupta et al.[9] aimed to develop a system that can classify messages as spam or non-spam. using dataset of 5,574 SMS messages with 747 spam and 4,827 non-spam messages divided into training and testing sets using different machine learning models such as Random Forest, Naive Bayes, SVM and Decision Tree. The results that SVM algorithm with TF-IDF feature extraction achieved 98.56% accuracy more than other datasets like recall, precision and F1-score.

Paras Sethi et al.[10] discusses the problem of SMS spam messages and the efforts to detect them using machine learning algorithms.it confirm the need for technical and legal measures to control the common abuse of mobile spam messages. This dataset is made of 5,574 classified SMS messages, where 4,827 messages are labeled as "ham" and 747 messages are labeled as "spam". The Naïve Bayes algorithm completed with a high check with 98.445%, the study also verified that the "Naïve bayes" algorithm outperforms the logistic regression and "random forest" algorithms and it's the least among the compared algorithms.

Houshmand et al[11] aims to classify SMS spams and compare their performance to find the best classifier for text messaging spam filtering by use machine learning techniques. the dataset uses 5574 text messages from the UCI Machine. Repository of learning compiled in 2012, comprising 425 spam messages extracted from the Grumbletext website, Each message in the dataset is labeled which is spam or ham, the distribution is 13.40% and 86.60%, respectively. It applies different machine learning models such as naive Bayes, SVM, k-nearest neighbor, random forests, and Adaboost. The best classifier is SVM with accuracy 97.64%. The next best classifier is naive Bayes with overall accuracy 97.50%.Based on the simulation, the most effective algorithms for SMS spam detection are multinomial naive Bayes with Laplace smoothing and SVM with a linear kernel..

Julis et al[12] It aims for filtering techniques to differentiate legitimate (Ham) than spam messages. This paper evaluates the effectiveness of several machine learning algorithms to filter SMS messages and improve the accuracy of spam detection. The dataset used in the experiment contains 2608 messages, 2408 collected from the SMS Spam Corpus and 200 collected manually. the labels for the dataset are "ham" for non-spam messages and "spam" for spam messages. The dataset used many machine learning algorithms based on training set, prediction time, and accuracy. The best accuracy score is Support Vector Machines (SVM) algorithm with 98%, the Naive Bayes algorithm showed the best prediction time, it provide effective spam and ham message collection in a short amount of time.

baqeel et al[13] aimed to develop a comprehensive approach to SMS spam detection using a mixed system of supervised and unsupervised machine learning techniques. The study was conducted to address the increasing challenge of SMS spam proliferation and the need for an effective spam detection system. The best algorithm identified in the research is Support Vector Machine(SVM) with an accuracy of 97%. The dataset used in the study consists of 5574 SMS messages obtained from the UCI machine learning repository.

kawade et al[14] discusses a solution in the research paper to lighten the economic impact of spam SMS on users and service providers, and the study concludes that machine learning techniques, specifically text classification, can effectively filter out spam SMS, with the NaiveBayes algorithm demonstrating the highest accuracy among other algorithms used in the study. The study utilized a dataset of 5568 SMS messages with attributes for class (Spam or Ham) and text content.

Almeida et al[15] addresses the economic cost and nuisance of spam emails, proposing new evaluation metrics and comparing algorithms to enhance content-based spam filtering, with the findings identifying the Linear Support Vector Machine as the most effective algorithm for automatic spam filtering across six well-known databases. The study was conducted across six well-known, large, and public databases, providing sufficient evidence to support the findings.

Jain et al[16]The goal of the research paper is to develop an effective method for detecting and filtering out smishing messages from spam messages to enhance mobile security. The paper was written to address the increasing threat of

smishing attacks and the limitations of existing security measures in combating this form of cyber-security threat. The proposed approach achieves the best accuracy using the Neural Network algorithm, and the research utilizes the SMS dataset v.1, which contains 5574 text messages in the English language, collected from various sources including the Grumbletext Website, NUS SMS Corpus, and SMS Spam Corpus v.0.1.

Syed Ali et al[17] consider the Random Forest algorithm to be the most accurate for detecting spam in SMS. They learned this using a database of 4,419 instances with 353 attributes, including 1,944 spam and 2,475 ham. They found that the most accurate algorithm was Random Forest, which reached 98.64 percent, followed by SVM and Naive Bayes, which were also OK.

Wael Hassan[18] chose a dataset consisting of 5574 messages containing 4827 ham and 747 spam. Many machine learning algorithms have been tried. They are tested on a dataset with the aim of achieving the most accurate algorithm. We tried DNN, LSTM: GRU, CNN, RCNN, HAN and RDML. The best accuracy obtained was 99.26 percent from RDML. The biggest advantage of RDML is that it reduces error rates and improves accuracy. The algorithms also have a high accuracy rate, but not like RDML.

Pumrapee Poomka et alblue[19] discuss spam prediction, they have by training the model based on the LSTM and GRU algorithms and comparing the effectiveness of the model based on machine learning . They used a data set consisting of 5,574 SMS. They are categorized into 747 spam messages and 4827 ham messages. They were split into 30 percent for training and 70 percent for testing. After testing, the results showed that there was no huge difference in the accuracy rate, as the LSTM accuracy rate was 98.18 percent and the GRU accuracy rate was 98.03 percent, a difference of less than 0.5 percent. They also compared the results with other algorithms such as Naive Bayes and SVM, and the results showed that the LSTM and GRU algorithms are more accurate than SVM and NB. Accordingly, they considered that the LSTM and GRU algorithms perform better.

In the context of combating spam messages, Odukoya Oluwatoyinblue[20] eliminated unnecessary data that did not affect the machine learning algorithm and selected factors based on consistency. They have a dataset containing 6,250 messages, including 1,562 spam messages and 4,688 ham messages. They experimented with four machine learning models: Multiplayer perceptron, Logistic Regression, Random Forest and Support vector machine. They tested these algorithms on a data set to compare them to arrive at the most accurate algorithm in detecting spam. The data was divided into a training set and a test set. To get the best results, all data was entered into the experiment. In comparison, the results showed the superiority of the Multilayer Perception algorithm by 96.89 percent. It also showed that the algorithm gave an increase of 3.03 percent in accuracy. Therefore, the Multilayer Perceptron algorithm has the best performance in detecting spam messages.

III. Proposed Methodology

Numerous algorithms are used, and a resarch is done on each algorithm before training the model using them on the datasets. The following diagram represents the steps the datasets went through to get the results.

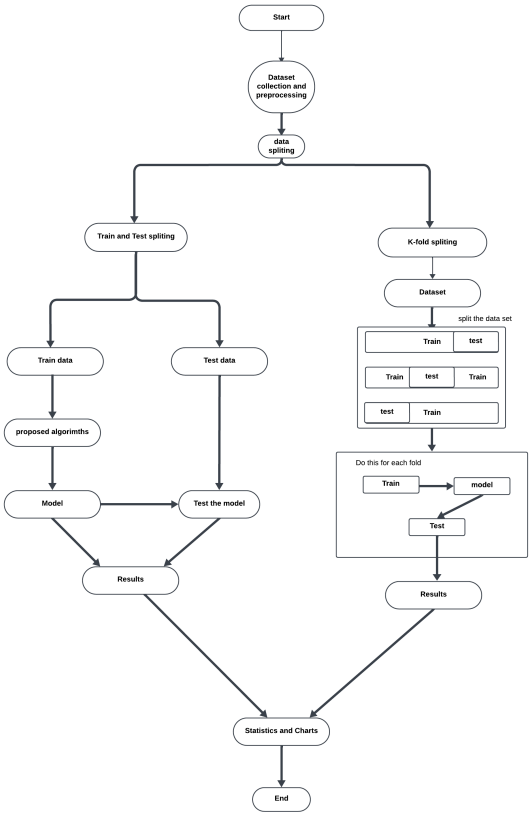

Fig. 1. SMS spam detection process

A. Datasets Descriptions

The dataset is a collection of text messages, used for spam detection. The messages are different in content, including regular conversations ("ham") and spam messages ("spam"). Each row provides a labeled example for a machine learning where the goal is to develop a model that can classify messages as spam or ham based on their content.

There are three dataset and they very similar to each other, each one consists of 1 feature which is text where the messages saved and has 1 label which is spam or ham, they just differ in text content and number of records.

All the datasets were divided into two partitions: for training 80% , and for testing 20%. The first dataset named

SMS Spam Detection Application(spam.csv) it has 5,572 records, the second dataset named **Spam SMS Classification(TrainDataset.csv)** and it has 4,457 records, and finally with the last dataset, that named **Filtering mobile phone spam(spamraw.csv)** which consists of 5,559 records and these messages were collected from people who were knowing that their contributions were going to be shared publicly available. **NOTE:** Click on datasets name for more information.

TABLE I
FEATURES OF DATASET 1

Feature	Type	Values
text	Classification	ham or spam

B. *Used Algorithms*

The mentioned datasets were trained to 5 different Machine Learning algorithms which were K Nearest Neighbor(kNN), SVM, Logistic Regression, Naive Bayes, and Decision Tree. For each of the algorithms there were performance metrics, these performance metrics were: The analysis compares the performance of five machine learning algorithms using metrics such as Accuracy, Precision, Recall, Mean Accuracy, and F1-score to determine their effectiveness in handling the given datasets.

1) **SVM:**
support Vector Machines(SVM) is a supervised machine learning algorithm used for classification and regression tasks. It excels in scenarios with high-dimensional data, where the number of features exceed the number of samples. The main objective of SVM is to identify a hyperplane that effectively separates data into distinct classes. The term "support vectors" denotes the data points nearest to the decision boundary, and SVM strives to maximize the difference, or the distance between the hyperplane and these support vectors. [21].

$$\begin{aligned} &\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i, \\ &\text{s.t. } y_i\left(w\cdot\varphi\left(x_i\right)+b\right)-1+\xi_i \geq 0, \\ &\xi_i \geq 0, \quad i=1,\ldots,l \end{aligned} \tag{1}$$

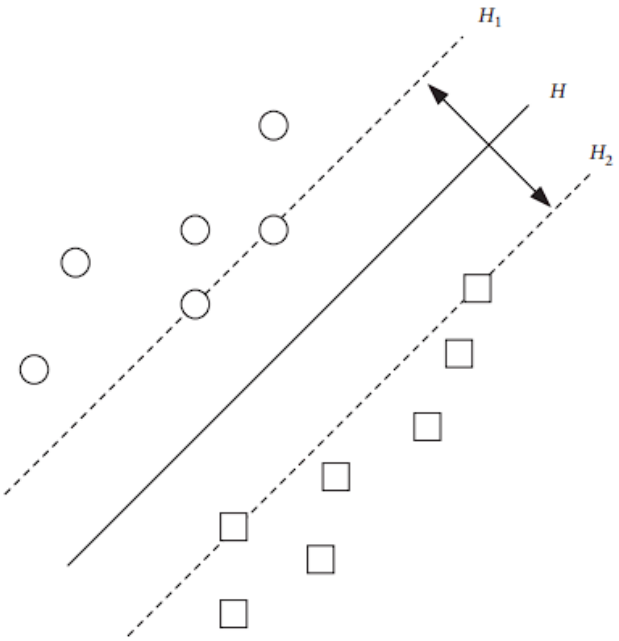

Fig. 2. the maximum margin hyperplane [1][21]

It changes the problems above into the following dual problem:

$$\begin{aligned} &\min_{\alpha} \frac{1}{2}\sum_{i=1}\sum_{j=1} y_i y_j \alpha_i \alpha_j K\left(x_i \cdot x_j\right) - \sum_{i=1}\alpha_i, \\ &\text{s.t. } \sum_{i=1}^{l}\alpha_i y_i = 0, \\ &0 \leq \alpha_i \leq C, \quad i=1,\ldots,l, \end{aligned} \tag{2}$$

in which i is Lagrange multiplier and K(xi · xj) (xi) · (xj) R is the kernel method, which represents the inner product of (xi) and (xj). If K(xi · xj) is positive definite or semidefinite,then we will use formula (2), formula(2) is a convex quadratic programming problem. And the final decision equation is as follows:

$$f_i = f\left(x_i\right) = \operatorname{sgn}\left(\sum_{j=1}^{l}\alpha_j y_j K\left(x_j, x_i\right) + b\right). \tag{3}$$

2) **Decision Tree:**
The supervised learning type includes the decision tree algorithm. Both classification and regression problems may be solved using them. Each node in the tree resemble to a class label, with attributes expressed on the tree's inner node. Any Boolean method with discrete characteristics may be described using the Decision tree. The entropy used to calculate the dataset randomness, it's value is always between that of 0 and 1. The value is the best when it is closer to 0. The classification of set S with respect to c is expressed in this equation "equation (1)". [22].

$$\text{Entropy}(S) = \sum_{i=1}^{c} \mathrm{P}_i \log 2^{\mathrm{P_i}} \tag{4}$$

Where P_i is a ratio of the sample from the subset and the $i-th$ attribute value.[22].
Information gain is a metric that intuitively informs how much knowledge of a random variable's value.

$$\text{Gain}(S, A) = \sum_{v \in V(A)} \frac{|S_v|}{|S|} \text{ Entropy } (S_v) \tag{5}$$

Where the range of A is $V(A)$, and S_V is a subset from set S equal to the attribute value of v [22].

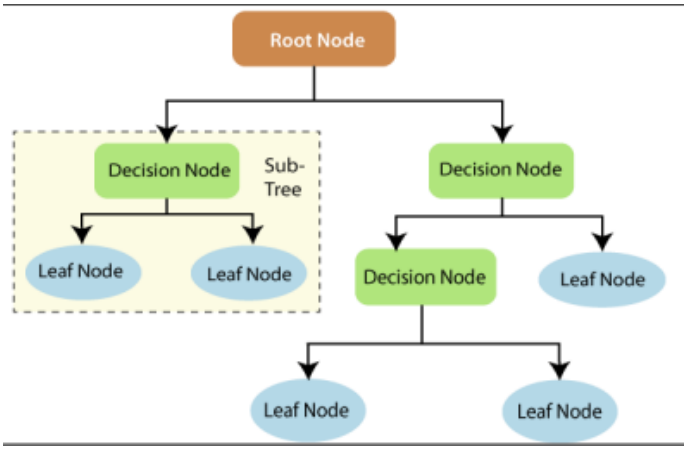

Fig. 3. Illustration of decision tree[22]

3) **Naïve Bayes:**
Naïve Bayes classifier is a supervised machine learning model, which is used in classification tasks, which include text classification. The program built on one of the bayes theorem, specifically the Multinomial Naïve Bayes that we used here to detect sms spam messages, it is a specific variant of the Naïve Bayes algorithm that is commonly used for text classification tasks, where the features represent the counts of words or other discrete data. Multinomial Naïve Bayes is computationally efficient and can handle large datasets with many features which makes it a practical choice for text classification tasks. [23]

Multinomial Naïve Bayes Classifier can be formulated as follows: A news article 'n' being of polarity 'p' is calculated as:

$$P(tk \mid p) = \frac{count(tk \mid p) + 1}{count(tp) + |V|} \tag{6}$$

where P(tk|p): represents the conditional probability that whether the term (tk) occurs in a news article of polarity p which is calculated as follows:

$$P(p \mid n) \alpha P(p) \pi_{1 \leq k \leq nd} P\left(tk \mid p\right) \tag{7}$$

Here, count (tk|p) means the number of times the term tk occurs in the news articles which have polarity (p) and count (tp) means the total number of tokens present in the news articles of polarity (p).[23]

4) **K – Nearest Neighbor:**
In this scenario, KNN(K-Nearest Neighbors) is a easy supervised machine-learning algorithm that can be used for classification and regression issues. This algorithm is 'non-parametric,' meaning it does not use a statistical model. Instead, it is based only on the training data (feature vectors and labels). This type of algorithm is known as a memory-based. The principle of KNN is about classification of data points based on the labels of the k nearest neighbor. The overall vote rule is applied to handle the situation where there is a tie between different classes. To calculate the distance between data points, the Euclidean distance is commonly used. Here is the Euclidean distance equation. [7]
.

$$\mathrm{d}(x, y) = \sqrt{(x-y) \cdot (x-y)} = \left(\sum_{i}^{m} (x_i - y_i)^2 \right)^{1/2}$$
$$x, y \in R^m \quad i = 1, 2, \ldots, m$$

5) **Logistic Regression:**
logistic regression is the process of modeling the probability of a discrete result given an input variable. A binary result, or something that can have two values, such as true or false, yes or no, and so on, is what most logistic regression equations represent. When there are more than two distinct discrete outcomes in a scenario, multinomial logistic regression can model the situation. A helpful analysis technique for classification issues is logistic regression, which is applied when attempting to ascertain which group a fresh sample most closely belongs into. Logistic regression is a helpful analytical tool since several parts of cyber security, such attack detection, are classification problems.
Another powerful supervised machine learning approach for binary classification issues is logistic regression (where target is categorical). Logistic regression can be best understood as a linear regression used to classification difficulties. In short, logistic regression models a binary output variable by applying the logistic function that is described below. The range of logistic regression is bounded between 0 and 1, which is the main distinction between it and linear regression. Furthermore, logistic regression does not require a linear relationship between the input and output variables, in contrast to linear regression. This results from the odds ratio being subjected to a nonlinear log transformation (which will be explained in a moment). [24]

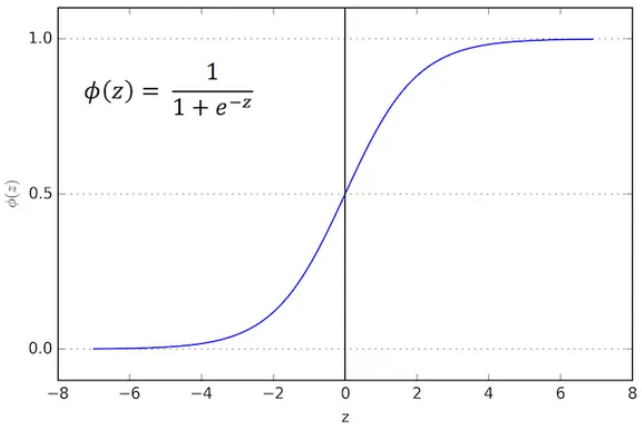

Fig. 4. "Spam Detection with Logistic Regression" by Natasha Sharma. [25]

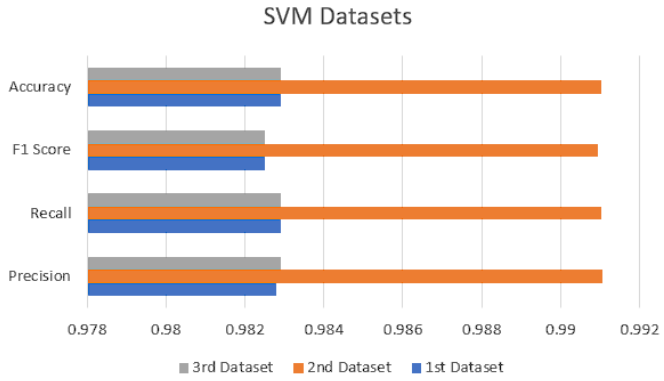

Fig. 5. All SVM datasets

C. *Performance Metrics*

Accuracy is the count of legitimately anticipated data from all the data. The count of accurately anticipated positives taken from the anticipated positives is the Precision Recall is the number of correctly anticipated positives from all the true positives. The number of accurately anticipated negatives out of all the expected negatives is known as specificity.

$$\text{Accuracy} = (\text{TN}+\text{TP})/(\text{TN}+\text{TP}+\text{FN}+\text{FP}) \quad (8)$$

$$\text{Precision} = \text{TP}/\text{TP}+\text{FP} \quad (9)$$

$$\text{Recall} = \text{TP}/\text{TP}+\text{FN} \quad (10)$$

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \quad (11)$$

IV. RESULTS

The results collected from Gradient Boosting, Naïve Bayes, Logistic Regression, Support Vector Machine(SVM), k-Nearest Neighbor(KNN), Decision Tree are shown below for each algorithm. All statistics of Algorithms with 80/20 data split.

1) SVM: After training and testing the datasets on SVM algorithm , these are the results for each dataset:

TABLE II
STATISTICS OF A LGORITHMS

SVM-Datasets	Precision	Recall	F1 Score	Accuracy
1st Dataset	0.9828	0.9829	0.9825	0.9829
2nd Dataset	0.9922	0.9921	0.992	0.9921
3rd Dataset	0.9829	0.9829	0.9825	0.9829

The provided datasets is related to spam detection, where the goal is to classify messages as either "ham" or "spam" based on their content. So, the best results of these dataset is Dataset 2 which has higher average accuracy with 98.73%, precision, recall, and F1 score compared to the other datasets. Therefore, Dataset 2 performs slightly better in terms of classification performance.

2) Naïve Bayes: After training and testing the datasets on Naïve Bayes algorithm, the two figures below show the results of using Naïve Bayes for each dataset:

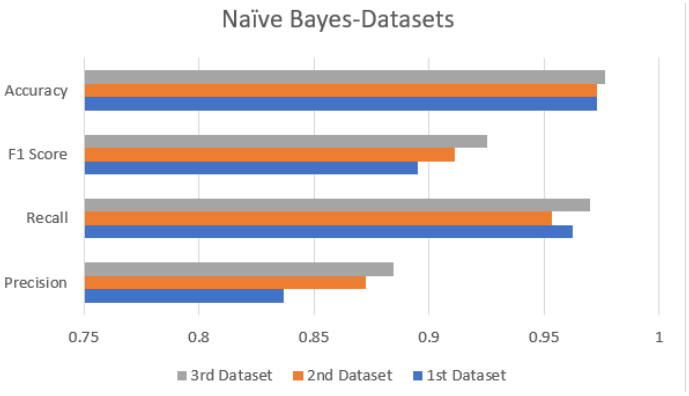

Fig. 6. All-datasets performance chart with data split

TABLE III
STATISTICS OF ALGORITHMS

Naïve Bayes-Datasets	Precision	Recall	F1 Score	Accuracy
1st Dataset	0.8366	0.9624	0.8951	0.973
2nd Dataset	0.8723	0.9535	0.9111	0.9731
3rd Dataset	0.8846	0.9699	0.9253	0.9767

The average accuracy of this classifier is approximately 97.43% and average precision is approximately 86.45% on the three datasets. This indicates the proportion of correctly classified instances for both spam and non-spam out of the total instances.

3) Decision tree:
After the spilting nd training of data using the Decision tree we reached these results :

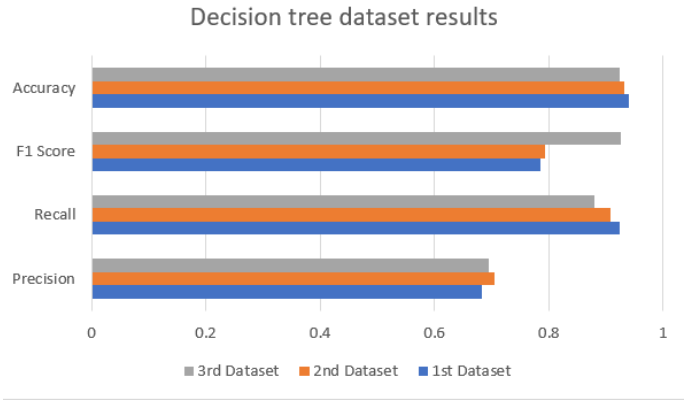

Fig. 7. All-datasets performance chart with Train test spilt

TABLE IV
STATISTICS OF ALGORITHMS

Datasets	Precision	Recall	F1 Score	Accuracy
1st Dataset	0.6833	0.9248	0.7859	0.9397
2nd Dataset	0.7048	0.907	0.7932	0.9316
3rd Dataset	0.6952	0.8795	0.9253	0.9247

The figure above shows the results of using the decision tree on the three data sets. It tells us that the average accuracy of this classifier is 93.2% and an average precision of 69.4%. The first dataset had the best accuracy and recall but also had the worst precision out of the three datasets.

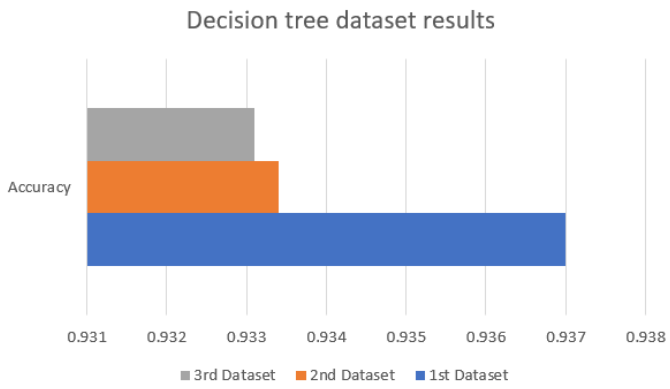

Fig. 8. All-datasets performance chart with k-fold

In the k-fold we notice that the first dataset has the highest mean accuracy out of the other datasets scoring a 93.7% mean accuracy with an increase of 0.5%.

4) KNN:
After the spilting nd training of data using the KNN we reached these results :

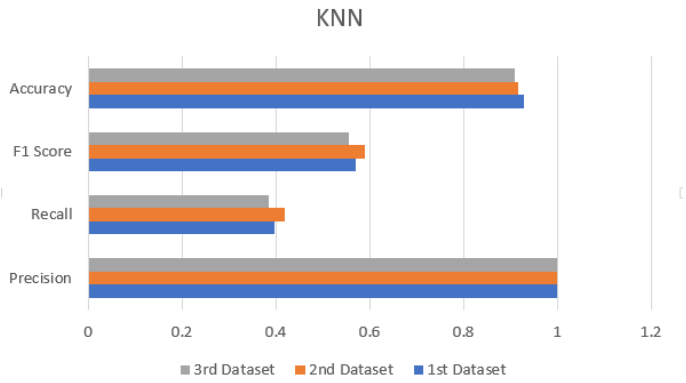

Fig. 9. All-datasets performance chart with Train test split

TABLE V
STATISTICS OF ALGORITHMS

Datasets	Precision	Recall	F1 Score	Accuracy
1st Dataset	1.000	0.3985	0.5699	0.9281
2nd Dataset	1.000	0.4186	0.5902	0.9159
3rd Dataset	1.000	0.3855	0.5565	0.9085

The figure above shows the results of using the KNN on the three data sets. It tells us that the average accuracy of this classifier is 91.75%. The first dataset had the best accuracy and recall but also had the worst precision out of the three datasets.

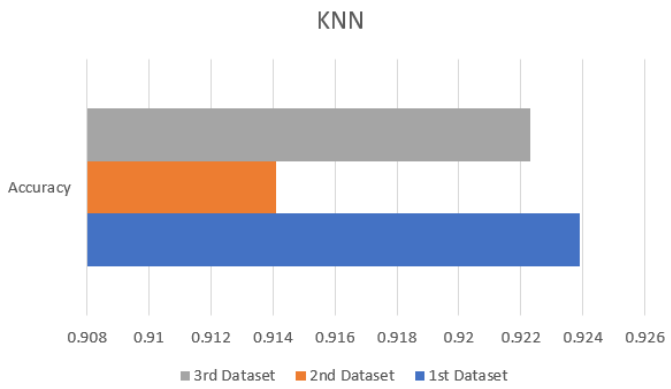

Fig. 10. All-datasets performance chart with Train test split

In the k-fold we notice that the first dataset has the highest mean accuracy out of the other datasets scoring a 92.39% mean accuracy with an decrease of 0.42%.

5) Logistic Regression:

After the spilting and training of data using the Logistic Regression we reached these results:

Fig. 11. All-datasets performance chart with Train test split

TABLE VI
STATISTICS OF ALGORITHMS

Datasets	Precision	Recall	F1 Score	Accuracy
1st Dataset	0.8944	0.9549	0.9236	0.9741
2nd Dataset	0.9370	0.9225	0.9297	0.9762
3rd Dataset	0.9128	0.9458	0.9290	0.9767

The figure above shows the results of using Logistic Regression across three spam datasets, with TrainDataset leading in accuracy (0.9762) and balanced performance. SpamRaw champions recall (0.9549), ideal for catching most spam even with slightly lower overall accuracy. Logistic Regression's average accuracy across the three datasets is 0.9756, meaning it correctly classifies nearly 98 out of every 100 emails on average.Choose TrainDataset for general spam classification, SpamRaw for prioritizing spam capture, and any for interpretability.

V. ANALYSIS

SVM consistently performs best across all datasets, offering high accuracy, recall, and F1-score. While KNN is the highest in precision for every dataset. Naive Bayes shines in recall, making it a good choice if capturing most spam emails is the top priority. Logistic Regression provides a balance between accuracy and interpretability, making it suitable when model transparency is important. KNN exhibits high precision for spam but suffers from low recall for ham emails, making it less suitable for general spam classification.

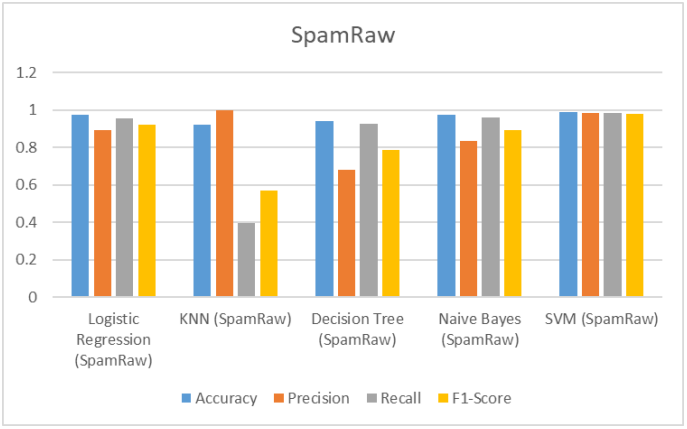

Fig. 12. SpamRaw

TABLE VII
FIRST DATA SET

Algorithms	Precision	Recall	F1 Score	Accuracy
svm	0.9828	0.9829	0.9825	0.9829
knn	1	0.3985	0.5699	0.9281
Naïve bayes	0.8366	0.9624	0.8951	0.973
Logistic regression	0.8944	0.9549	0.9236	0.9741
Decision tree	0.6833	0.9248	0.7859	0.9397

SpamRaw: in the SpamRaw dataset, SVM reigns supreme, boasting the highest accuracy (0.9914) and F1-score (0.9825). While KNN is the highest in precision (1) Best Algorithm: SVM (SpamRaw) for absolute accuracy and balanced performance.

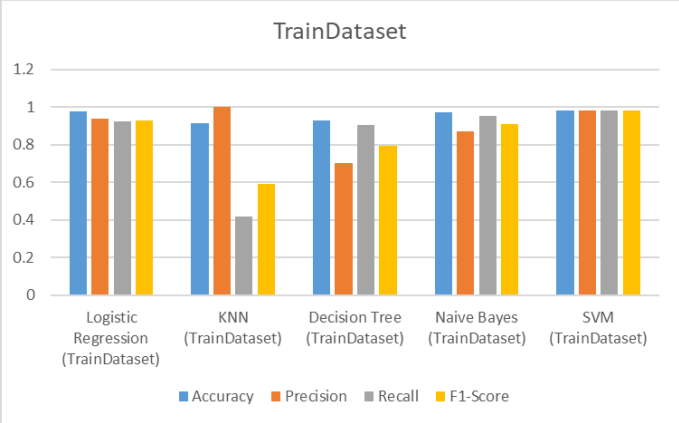

Fig. 13. TrainDataset

TABLE VIII
SECOND DATA SET

Algorithms	Precision	Recall	F1 Score	Accuracy
svm	0.9922	0.9921	0.992	0.9921
knn	1.000	0.4186	0.5902	0.9159
Naïve bayes	0.8723	0.9535	0.9111	0.9731
Logistic regression	0.9370	0.9225	0.9297	0.9762
Decision tree	0.7048	0.907	0.7932	0.9316

In the TrainDataset, SVM remains dominant with high accuracy (0.9829), F1-score (0.9826).While KNN is the highest in precision(1). Best Algorithm: SVM (TrainDataset) for overall performance and accuracy.

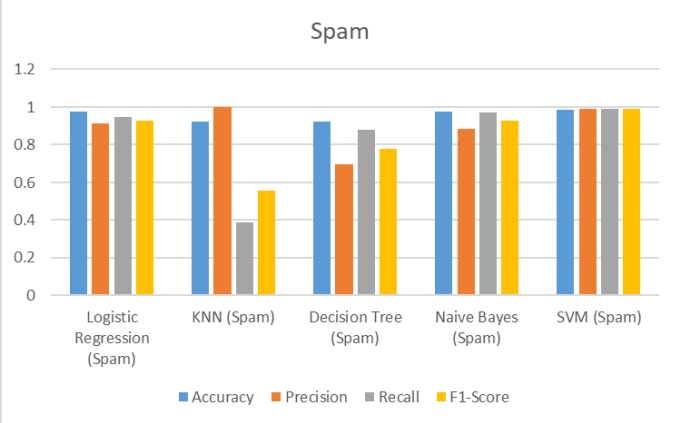

Fig. 14. Spam

TABLE IX
THIRD DATA SET

Algorithms	Precision	Recall	F1 Score	Accuracy
svm	0.9829	0.9829	0.9825	0.9829
knn	1.000	0.3855	0.5565	0.9085
Naïve bayes	0.8846	0.9699	0.9253	0.9767
Logistic regression	0.9370	0.9225	0.9297	0.9762
Decision tree	0.6952	0.8795	0.9253	0.9247

Spam: in the Spam dataset, SVM again emerges as the champion, achieving the highest accuracy (0.9829), F1-score (0.9829). While KNN is the highest in precision (1). Best Algorithm: SVM (Spam) for absolute performance across all metrics. Naive Bayes (Spam) for high recall (0.9560) if capturing most spam is essential.

VI. Conclusion

Machine learning in detection of sms spam messages has proven to be a effective approach and formidable in addressing the growing menace of non-useful text messages. Through the utilization of various algorithms and features the classification accuracy can increase even more. Support Vector Machines(SVM) surpassed in detecting sms spam messages in most datasets with average accuracy of 98,57%, which indicates that it is the best algoritm for sms spam detection. Machine Learning can be applied to many fields, not just in spam detection, it can be used to predict anything from stock prices to results of sports matches using analysis information and datasets to train on it and also can come up with solutions for big problems or fixing any sitiuation, which makes it a very useful tool for humanity. And this tool will only keep improving and producing better results.

VII. Acknowledgment

At the end, Let's take a moment to express our deepest gratitude to the dedicated team at Misr International University, particularly the faculty of computer science, for their relentless efforts in elevating the standards of this esteemed institution. Without their unwavering commitment, the university wouldn't be the thriving hub of knowledge and learning that it is today. A heartfelt appreciation goes to none other than Prof. Mohamed Shebl El Komy, the University President, for his exceptional leadership and commitment towards creating a conducive learning environment, we extend our sincere thanks to Associate Prof. Ayman Nabil dean faculty of Computer Science, Prof. AbdelNasser Zaied Vice Dean of Student Affairs and Professor of Computer Engineering for giving us the chance to learn in this virtuous university and running it efficiently. And finally, Our journey in the field of artificial intelligence would not have been the same without the invaluable guidance and instruction of Dr. Diaa AbdelMoneim associate professor in information systems and our instructor in artificial intelligence course, and also Eng. Tarek Talaat teaching assistant for their continued guidance and supporting us in our work.

References

[1] T. A. Almeida, J. M. G. Hidalgo, and A. Yamakami, "Contributions to the study of sms spam filtering: new collection and results," in *Proceedings of the 11th ACM symposium on Document engineering*, 2011, pp. 259–262.

[2] L. GuangJun, S. Nazir, H. U. Khan, and A. U. Haq, "Spam detection approach for secure mobile message communication using machine learning algorithms," *Security and Communication Networks*, vol. 2020, pp. 1–6, 2020.

[3] A. Alzahrani and D. B. Rawat, "Comparative study of machine learning algorithms for sms spam detection," in *2019 SoutheastCon*. IEEE, 2019, pp. 1–6.

[4] P. Navaney, G. Dubey, and A. Rana, "Sms spam filtering using supervised machine learning algorithms," in *2018 8th international conference on cloud computing, data science & engineering (confluence)*. IEEE, 2018, pp. 43–48.

[5] O. Agboola, "Spam detection using machine learning and deep learning," Ph.D. dissertation, Louisiana State University and Agricultural & Mechanical College, 2022.

[6] M. Gupta, A. Bakliwal, S. Agarwal, and P. Mehndiratta, "A comparative study of spam sms detection using machine learning classifiers," in *2018 eleventh international conference on contemporary computing (IC3)*. IEEE, 2018, pp. 1–7.

[7] F. Z. El Hlouli, J. Riffi, M. A. Mahraz, A. El Yahyaouy, and H. Tairi, "Detection of sms spam using machine-learning algorithms," in *Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco*. Springer, 2020, pp. 429–440.

[8] Y. Kontsewaya, E. Antonov, and A. Artamonov, "Evaluating the effectiveness of machine learning methods for spam detection," *Procedia Computer Science*, vol. 190, pp. 479–486, 2021.

[9] S. D. Gupta, S. Saha, and S. K. Das, "Sms spam detection using machine learning," in *Journal of Physics: Conference Series*, vol. 1797, no. 1. IOP Publishing, 2021, p. 012017.

[10] P. Sethi, V. Bhandari, and B. Kohli, "Sms spam detection and comparison of various machine learning algorithms," in *2017 international conference on computing and communication technologies for smart nation (IC3TSN)*. IEEE, 2017, pp. 28–31.

[11] H. Shirani-Mehr, "Sms spam detection using machine learning approach," *unpublished) http://cs229. stanford. edu/proj2013/Shir aniMeh r-SMSSpamDetectionUsingMachineLearningApproach. pdf*, 2013.

[12] M. R. Julis and S. Alagesan, "Spam detection in sms using machine learning through textmining," *Internatıonal Journal Of Scıentıfic & Technology Research*, vol. 9, no. 02, 2020.

[13] H. Baaqeel and R. Zagrouba, "Hybrid sms spam filtering system using machine learning techniques," *2020 International Conference on Computer and Information Sciences (ICCIS)*, pp. 1–6, 2020.

[14] D. R. Kawade and K. S. Oza, "Content-based sms spam filtering using machine learning technique," *International Journal of Computer Engineering and Applications*, vol. XII, pp. 626–630, 2018.

[15] T. A. Almeida and A. Yamakami, "Content-based spam filtering," in *Proceedings of the [Conference Name]*, 2010, pp. 1–7.

[16] S. S. Ali and J. Maqsood, ". net library for sms spam detection using machine learning: A cross platform solution," in *2018 15th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*. IEEE, 2018, pp. 470–476.

[17] A. K. Jain, S. K. Yadaf, and N. Choudhary, "A novel approach to detect spam and smishing sms using machine learning techniques," *International Journal of Computer Engineering and Applications*, vol. XII, no. IV, pp. 626–630, 2020.

[18] W. H. Gomaa, "The impact of deep learning techniques on sms spam filtering," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, 2020.

[19] P. Poomka, W. Pongsena, N. Kerdprasop, and K. Kerdprasop, "Sms spam detection based on long short-term memory and gated recurrent unit," *International Journal of Future Computer and Communication*, vol. 8, no. 1, pp. 11–15, 2019.

[20] O. Oluwatoyin, A. Bodunde, G. Titus, and A. Ganiyu, "An improved machine learning-based short message service spam detection system," *International Journal of Computer Network and Information Security*, vol. 10, no. 12, p. 40, 2019.

[21] W. Xie, Y. She, and Q. Guo, "Research on multiple classification based on improved svm algorithm for balanced binary decision tree," *Scientific Programming*, vol. 2021, pp. 1–11, 2021.

[22] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.

[23] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and bernoulli naïve bayes for text classification," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*. IEEE, 2019, pp. 593–596.

[24] T. W. Edgar and D. O. Manz. (2017) Logistic regression. Logistic regression is a classification method that can be used exclusively in two-class problems.

[25] N. Sharma, “Spam detection with logistic regression,” *Towards Data Science*, may 2018, logistic Regression equation.