An Empirical study of Short Messaging Service - Smishing Detection using Machine Learning Algorithms

A. Asirvatham[1], C. Meenakshi[2]
Research Scholar[1], Associate Professor[2]
Department of Computer Science, Vels Institute of Science Technology and Advanced Studies, Pallavaram, Chennai, India[1,2]
anishaasirvatham@outlook.com[1]
cmeenakshi.scs@velsuniv.ac.in[2]

Abstract

Phishing attack with SMS basically, smishing occurs when a SMS message which looks legitimate makes the common make believe that it's a the genuine message to open and click the links which may be malicious. Online surveys prove that phishing attack as the leading attack for web content which targets the financial concerns. Smishing is a way to detect the SMS phishing techniques. Now-a-days, Smishing is a matter of concern which creates havoc among the common people. The individuals are not aware of the phishing techniques and they fall prey to the fraudster messages and shell out a huge sum from their savings. Also, the awareness about the same is given by the government and many social media, to the people, but then too these kinds of frauds are still happening. This paper shows the empirical study on the detection of smishing attacks using machine learning algorithms like SVM, Random forest and their ensemble models. The most appealing result is achieved by the ensemble model with Support Vector Machine and Random forest using voting classifier with accuracy of 99.01% implementing with the SMS UCI machine learning repository dataset.

Keywords—Smishing, Random forest, SVM, Ensemble, machine learning

I. Introduction

SMS spam detection has been a significant area of research due to the proliferation of unsolicited messages affecting mobile users worldwide. Various methodologies, ranging from machine learning techniques with ensemble models, have been proposed to enhance the accuracy and reliability of spam detection systems. As, Li et al. (2024) developed SpamDam, a framework addressing challenges like the scarcity of public SMS spam datasets and privacy concerns. It uses blockchain technology to collect labeled SMS data securely and employs machine learning models to detect the spam content in the received messages. The Naive Bayes classifier has been widely used for spam detection due to its reliable, simplicity and effectiveness. By calculating the probability of a message being spam based on the presence of certain words, it provides an empirical data to filter the unwanted messages. Gedam and Banchhor (2023) proposed an enhanced SMS spam detection framework that integrates blockchain technology with machine learning. This approach ensures transparent and secure data collection, which is then used to train models incorporating natural language processing and ensemble methods, thereby improving classification accuracy. It explored the use of ensemble learning techniques, combining Support Vector Machine (SVM), Naive Bayes, Extra Trees, and a Voting Classifier.

This ensemble approach achieved a 94% accuracy rate in detecting spam messages, highlighting the effectiveness of combining multiple classifiers. Uddin et al. (2024) investigated transformer-based language models, such as RoBERTa, for SMS spam detection. Their fine-tuned model achieved a high accuracy of 99.84%. They also incorporated Explainable Artificial Intelligence (XAI) techniques to enhance model transparency in the classification process. Gadde et al. (2021) compared various machine learning and deep learning techniques for SMS spam detection. It demonstrated that deep learning models, particularly those employing neural networks, offer superior performance in identifying spam messages. Saleh et al. (2019) proposed an intelligent spam detection model inspired by artificial immune systems. This bio-inspired approach enhances the system's ability to recognize and filter out spam messages effectively.

II. Review of Literature

Ulus et al. (2022) explored transfer learning using augmentation and stacking for SMS spam detection. Their approach demonstrated that transfer learning could effectively enhance spam detection performance, especially when labelled data is limited. Silpa et al. (2023) developed a meta-classifier model combining Multinomial Naive Bayes and Linear Support Vector Classifier algorithms. This hybrid model improved the accuracy of SMS spam detection, showcasing the benefits of integrating multiple algorithms. Ahmed et al. (2022) analysed machine learning techniques for spam detection in both email and IoT platforms. Their study highlighted the challenges and research opportunities in extending spam detection methods to IoT environments.

Bari et al. (2023) utilized natural language processing and machine learning for classifying SMS and email spam. Their approach emphasizes the importance of text pre-processing and feature extraction in improving detection accuracy. Reaves et al. (2016) investigated the detection of SMS spam in the context of legitimate bulk messaging. Their study addressed the challenges posed by adversaries attempting to bypass spam filters, contributing to the development of more robust detection systems. D. A. Oyeyemi et al. (2024) explains the effectiveness of combining advanced NLP models with BERT and traditional machine learning classifiers to enhance SMS spam detection. The approach not only improves accuracy but also reduces false positives, to protect users from spam and phishing attacks. G. Sousa et al. (2021) proposed a method to demonstrate competitive performance, achieving high accuracy in finding spam and legitimate messages. The combination of skip-gram embedding's with shallow networks proves effective, offering a balance between computational efficiency and classification performance. The approach is particularly for applications where computational resources are limited

There are prominent challenges on SMS Spam Detection like Spammers adapt the strategies to bypass the existing spam detection system, finding the large datasets for training is challenging which gives a biased model to detect the spam and ham messages. While checking for spam and ham messages, collection of personal information may raise the privacy and ethical issues.

SMS spam detection has evolved from simple keyword-based filters to sophisticated models employing machine learning, deep learning, and block chain technologies. While significant progress has been made, ongoing research is essential to address emerging challenges and ensure the security and privacy of mobile communications. SMS spam detection can be performed using the following feature extraction methods like the following:

- Bag of Words (BoW) Representation: This finds the single word or pairs/triplets of words. It also counts, how many times, each word appears in a text message.
- TF-IDF (Term Frequency - Inverse Document Frequency): It Measures the importance of a word in the context of text message available. It also combines the term frequency and inverse document frequency across the document.
- Presence of Specific Keywords: Spam related words like free, cash, win, offer, etc. are used in the spam messages.
- Message Length: It's the number of characters or words with specific pattern
- Capitalization and Punctuation: Normally the spam messages use capital letters with use of exclamation marks or question marks
- Presence of URLs and Email Addresses: Spam messages contains some links to external websites and also the presence of email addresses with spelling mistakes is also spam.
- Character Patterns: spam messages with repeated characters, detection of misspellings or altered words with combination of characters and numbers like fr33 instead of free.
- Language Features: here, the grammar and syntax plays a vital role in the framing of spam messages. In legitimate messages, the grammar will be proper and the way of expressing the thoughts will be clear unlike spam messages.

Machine Learning Algorithms used in this study:

A. Support Vector Machine – Applied for both classification and regression, Support Vector Machines (SVM) are a supervised learning tool. SVM is very good for classification tasks even if it can control regression problems. SVM looks for the most appropriate hyper plane in an N-dimensional space that separates data points into several groups. The method maximizes the interval between the closest points of different classes. Support Vector Machines, or SVM, A fundamental method in machine learning applied for classification problems is SVM. It provides a means to find a hyper plane in an n-dimensional space separating the data points into two groups, one on either side of the plane. The variety of features determines the hyper plane. The hyper plane is represented by a line in two dimensions. A hyper plane is just a 2D plane in 3D space. In order to optimize the distance between the data points, SVM looks for a hyper plane. Support vectors are the names given to these points in space. Support Vector Machines are the name given to this algorithm for that reason. The margins are the distances between the support vectors and the hyper plane. There are two types of SVM. They are as follows:

- Linear: The Linear SVM is a type of SVM in which a line can be used to split the support vectors linearly.
- Nonlinear: An SVM in which a line cannot linearly separate the support vectors is known as a non-linear SVM. The resulting hyper plane may have the shape of a circle, hyperbola, etc. Kernels can be used for non-linear situations. A particular kind of function called a kernel is used to transform lower-dimensional problems into higher-dimensional ones since doing so makes separation easier.

B. Random Forest – In order to produce predictions, we endorse on all the trees using the Random Forest algorithm, a potent tree learning method in machine learning. They are frequently employed for tasks involving regression and classification. This kind of classifier makes predictions by utilizing a large number of decision trees. In random forest, random segments are trained in the tree to get the average value, which gives the increased accuracy. This model is widely used in ensemble models. It handles Missing Data: During training, missing values are handled automatically, doing away with the need for human imputation. The algorithm provides useful insights for feature selection and interpretability by ranking features according to their significance in making predictions. Adapts well to complex and large data sets without experiencing appreciable performance reduction. Because of its versatility, algorithms can be used for both regression problems (like predicting continuous values) and classification tasks (like predicting categories).

- Ensemble SVM + RF – Ensemble models are the techniques in machine learning that combines multiple individual models to improve the predictive performance to give the highest accuracy of the existing models. Two base algorithms used in ensemble learning are Support

979-8-3315-9610-1/25/$31.00 ©2025 IEEE

Vector Machines (SVMs) and Random Forest (RF)

III. METHODOLOGY

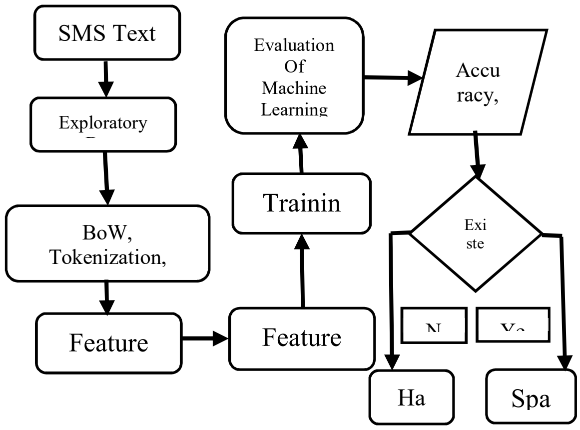

Figure 1. – Methodology Diagram

The above Figure1 shows the SMS dataset in which the exploratory analysis is performed. Then the data processing is performed by removal of stop words, Bag of words which counts, how many times each word appears in the text message, Tokenization where the small words are taken as tokens, stemming leads to get the root words from the word. Taking the training to testing ratio to 80:20. Then evaluating the different machine learning algorithm like SVM, RF and Ensemble SVM+RF. The performance metrics like accuracy. Precision, Recall and F-Score. We are checking the existence of spam words in the SMS for ham messages or spam messages.

This model shows the appropriate means to detect the SMS as spam or ham messages among the n number of messages received from a single source to dupe the common people. The exploratory data analysis is performed before the application of pre-processing techniques like Bag of Words (BoW), Tokenisation of words, selection of vectoriser, then applying the proposed ensemble SVM and RF model. The model is optimised further to get the improved accuracy and precision values. The above figure shows the flowchart of processes performed to detect the spam or ham messages for the proposed model.

Dataset- The dataset used for this article is from the Kaggle and the different sources like GrumbleText Website, NUS SMS Corpus(NSC) and from Caroline Tag's PhD Thesis. The SMS spam dataset collection by UCI Machine Learning dataset is consisting of Ham messages which is considered to be legitimate messages from a genuine source and is not represented or considered as a spam message.

Table 1: Dataset showing the total number of messages.

Total Number of Messages	Ham Messages	Spam Messages
5572	4825	747

Table 1 shows the messages from dataset. It consists of 4825 messages as Ham messages and 747 messages as Spam messages. This dataset has 5574 unique values having 86.6% ham message and 13.4% Spam message

.

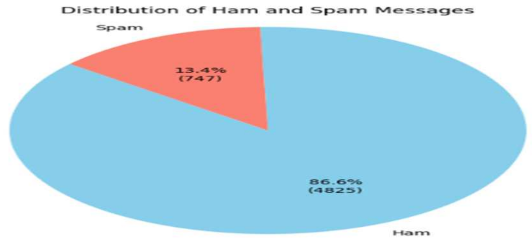

Figure 2. Distribution of Spam and Ham messages

The pie chart showing the distribution of Spam and Ham message in shown in Figure.2. This figure shows that 13.4% messages are spam messages and 86.6% messages are ham messages.

A. Data Pre-Processing:

SMS Dataset uses the Exploratory data analysis methodology to obtain a preliminary information about the data. Text pre-processing technique consists of Bag of Words, Tokenization of text means converting the big sentences into small texts which are meaningful, checking for the presence of specific keywords like Congratulations, Winner, etc., Finding the length of the message as readable and understandable to the reader, Eliminate stop words and punctuations from text which helps to simplify the data and reduce the size of the data to speed up the processes, Stemming is performed to reduce the word to its base word to normalize the text. Presence of valid URL's and email addresses, check the writing style to find the pattern of language. Also, to reduce the repetition of unwanted characters in the message. The dataset is transformed to identify as spam and ham messages.

B. Word Embedding Technique

Text data must first be converted into a numerical format for machine learning algorithms that can only work with numerical data.

C. TF-IDF Vectorizer -

The TF-IDF Vectorizer measures the significance of a words in sets in a document. It calculates a score for each word in a document based on the frequency of words in the document as term frequency, and the inverse calculation is the inverse document frequency. The commonly appearing words have higher term frequency score, while words that are less common have the higher inverse document frequency score. The TF-IDF vectorizer forms a matrix where each row signifies a document, each column represents a word, and the numeric values indicate the relative importance of each word in the corresponding document.

$$TF = \frac{\text{Frequency of word in a document}}{\text{Total number of words in that document}} \quad [1]$$

$$DF = \frac{\text{Documents containing word W}}{\text{Total number of documents}} \quad [2]$$

$$IDF = \log\left(\frac{\text{Total quantity of document}}{\text{Documents containing word W}}\right) \quad [3]$$

$$TF - IDF = TF \times IDF \quad [4]$$

D. CountVectoriser – This gets the count of token that is the count of words. This consists of the collection of documents with the count of words matrix.

E. Classification Model in Selection of Vectorizer with Machine Learning Techniques:

The model used in this article is Support Vector Machine (SVM), Random Forest (RF) and Ensemble (SVM + RF) to detect the spam contents in the SMS. It can also help to create a recommendation system to detect the spam contents in the message and text data mining. It uses the vectorizers to convert the text in the dataset into numerical vectors and applies to the mentioned machine learning algorithms. The figure 3 below, shows the frequency of single and two words existence in the spam messages as N-grams. This shows a common word “to” appeared 2252 times, “you” appeared 2245 times and so on.

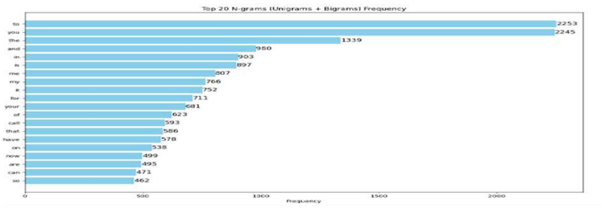

Figure 3. N-grams frequency of frequently used words

Algorithm

TRAINING PHASE

Step 1. Input: The Labelled Dataset with the SMS messages having both the spam and non-spam text messages

Step 2. Perform Exploratory Data Analysis

Step 3. Apply data pre-processing techniques like:

- Bag of Words representation
- Tokenize the given text
- Presence of specific keywords
- Message Length
- Eliminate stop words and punctuations
- Presence of URL's and email addresses.
- Language pattern
- Repetition of characters
- Stemming

Step 4. Perform TF-IDF(Term Frequency - Inverse Document Frequency) vectorization and use of Count vectorizer for feature extraction

Step 5. Split data into training and testing sets

Step 6. Train with machine learning model for Proposed Ensemble SVM+RF, then evaluate the model performance with accuracy and precision.

Step 7. Save the model.

Step 8. Output: The model is trained for Spam detection mechanism.

Step 9. Stop

The above given algorithm is the proposed model to detect the spam SMS messages as ham or spam messages. The algorithm is performed in 2 phases like Training Phases and Testing Phase. In Training phase, the message is taken from the mentioned dataset. The exploratory data analysis to get the mean and median statistical values as 80 and 61, respectively. Then different preprocessing techniques are performed to clean the raw data as, BoW, Tokenization, removal of stop words, presence of specific words, presence of URL’s and emails, repetition of characters and stemming. The mentioned preprocessing with count vectorizer and TFIDF vectorizers are performed for extraction of features. The given dataset is splitted into 80% training set and 20% testing set. The proposed Ensemble SVM+RF is performed on the training set to achieve the performance metrics with accuracy as 99.01%.

TESTING PHASE

Step 1. Input: The Labelled Dataset with the SMS messages having both the spam and non-spam text messages

Step 2. Preprocess the given SMS text using the techniques from Training Phase.

Step 3. Pre-processing using TF-IDF and count vectoriser to vectorise the text

Step 4. Load the pre-trained model with SVM using CalibratedClassifierCV, random forest classifier and ensemble model

Step 5. Use voting classifier to classify the text messages using soft voting with probability output

Step 6. Output: Classification reports with performance metrics like Accuracy, Precision, Recall and F-Score.

Step 7. Stop

In Testing phase, the message is taken from the mentioned dataset. Then different preprocessing techniques are performed to clean the raw data as mentioned in the training phase. The mentioned preprocessing with count vectorizer and TFIDF vectorizers are performed for extraction of features. The given dataset is separated into 80% training set and 20% testing set. The proposed model with Ensemble SVM + RF using voting classifier is performed along with the simple Support Vector Machine and Random forest models are performed. SVM,CalibratedClassifierCV is used to perform probability calibration with the given SMS UCI machine learning repository dataset.

IV. Results and Findings

A. Metrics Evaluation

The machine learning algorithms were measured based on its accuracy, precision, recall, and F1 scores, but only the accuracy metric will be used to evaluate the model with the best performance.

- Accuracy- It is defined as the ratio of addition of true positive and true negative to the sum of all positives and negatives.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad \ldots.[5]$$

- Precision- It is defined as the ratio of true positive to the sum of true positive and false positive.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \ldots.[6]$$

1. Recall- It is defined as the ratio of true positive to the sum of true positive and false negative.

$$\text{Recall} = \frac{TP}{TP + FN} \quad ....[7]$$

2. FI score-

$$\text{F1 Score} = \left(\frac{2*(Precision*Recall)}{Precison+Recall}\right) \quad [8]$$

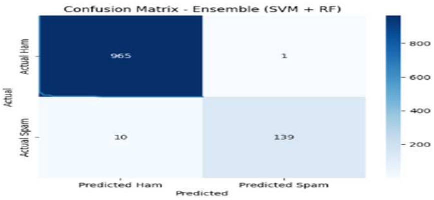

Figure 4 – Confusion matrix of the proposed system (Ensemble SVM +RF)

Figure 4. Shows the confusion matrix of the proposed system ensemble SVM and RF. This shows the actual ham to predict ham messages as 965 which is correctly predicted. There are 139 SMS messages which is true negative value mentioning that the message in spam. Stating the true negative values to predict the spam messages from the same dataset. These values are considered with the testing set of 20% as 1115 messages. Confusion Matrix is used as a performance metrics to evaluate the machine learning models. It shows the detailed information about the correct and incorrect predictions in comparison with the actual outcomes from the considered dataset.

Table 2: Summary of accuracy of the machine learning algorithms.

Sr.No	Algorithm	Accuracy
1.	Support Vector Machine	0.9892
2.	Random Forest	0.9793
3.	Proposed Ensemble (SVM + RF)	0.9901

The above table 2, shows the summary of the different machine algorithms Support Vector Machine (SVM), Random forest (RF) and Ensemble SVM + RF. Comparing the accuracy with the base machine learning models with SVM, RF and ensemble models, the SVM gives 98.92%,RF shows 97.93 and the proposed Ensemble model shows 99.01% respectively.

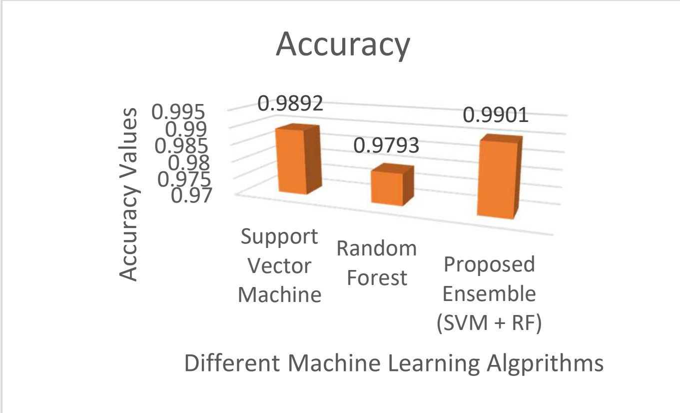

Figure 5. Accuracies of different Machine learning Algorithm

The Precision, Recall and FI-Score of the spam messages with Random forest model is 1, 0.85 and 0.92 respectively. The same values of the spam messages with SVM model is 0.97, 0.95 and 0.96 respectively. The values of the spam messages with Ensemble model is 0.99, 0.93 and 0.96 respectively.

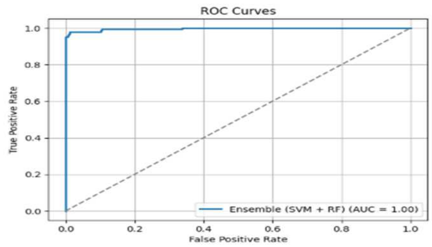

Figure.8. ROC of Ensemble (SVM + RF) model

The above Figure 8 shows the ROC-AUC curve with the proposed ensemble model. The values of the precision, recall and f-score values of spam messages with Ensemble model is 0.99, 0.93 and 0.96 respectively. The Curve is slightly tilted to achieve the AUC as 100%.

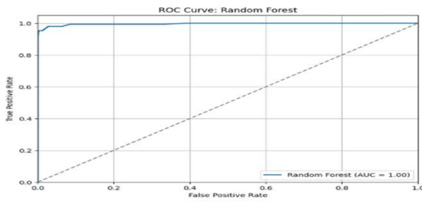

Figure.9. ROC of Random Forest model

The above Figure 9 shows the ROC-AUC curve with the Random forest model applied in the same dataset. The Precision, Recall and FI-Score of the spam messages with Random forest model is 1, 0.85 and 0.92 respectively. The Curve is slightly tilted to achieve the AUC as 100%.

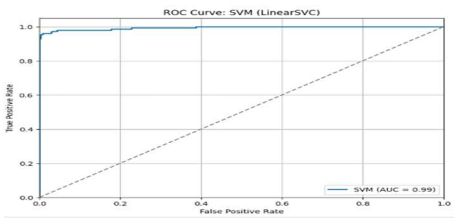

Figure.10. ROC of Support Vector Machine model

The above Figure 10 shows the ROC-AUC curve with the Support Vector Machine model applied in the same dataset. The Precision, Recall and FI-Score of the spam messages with SVM model is 0.97, 0.95 and 0.96 respectively. The Curve is slightly tilted to achieve the AUC as 99%. The ROC-AUC curve of the study performed to detect the spam SMS messages from the total 5572 messages available in the given dataset. The figure shows that Random forest and proposed Ensemble SVM+RF gives 100% results and the ROC-AUC of Support vector machine shows 99%.

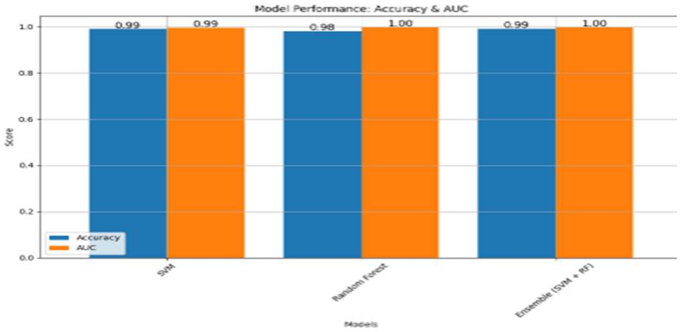

Figure 11. Performance Accuracy and AUC of SVM,RF and ensemble SVM + RF models

Figure 11, shows the performance metrics with the accuracy and AUC of SVM, RF and Ensemble SVM + RF is shown below in graph. SVM shows the accuracy of 98.92% and AUC of 99.00 %, Random forest shows the accuracy of 97.93% and AUC with 100%. The Ensemble SVM + RF shows the accuracy of 99.01% and AUC as 100%. The Precision of all 3 models are 99% for ham messages, SVM models with 97%, RF with 100% and Ensemble model with 99%. The ROC Curve (Receiver Operating Characteristic) curve visualizes the performance of classification model at various settings. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR). The area under the ROC curve (AUC-ROC) provides a single metric summarizing the model's overall ability to distinguish between classes, with higher AUC indicating better performance

V. Conclusion

In this paper, the machine learning algorithms like Support Vector Machine, Random forest and ensemble models are studied, where an ensemble model with SVM and RF is proposed to detect the SMS message as spam or ham message. The dataset used is SMS spam detection from UCI machine learning repository. The well know word embedding technique using TF-IDF and count vectoriser is implemented in the proposed system. The comparison was done with machine learning algorithms like Support Vector Machine, Random forest and ensemble SVM+RF models. Various parameters with the whole dataset without splitting gives the data as 4825 as Ham messages and 747 as Spam messages among the 5572 messages present in the dataset. Hence, we conclude that the ensemble model with SVM and RF helps to detect the SMS as spam or ham message with 99.01 % accuracy which outperforms the SVM and RF machine learning models.

V. References

[1] Y. Li, R. Zhang, W. Rong, and X. Mi, "SpamDam: Towards Privacy-Preserving and Adversary-Resistant SMS Spam Detection," arXiv preprint arXiv:2404.09481, 2024. [Online]. Available: https://arxiv.org/abs/2404.09481

[2] D. A. Oyeyemi and A. K. Ojo, "SMS Spam Detection and Classification to Combat Abuse in Telephone Networks Using Natural Language Processing," arXiv preprint arXiv:2406.06578, 2024. [Online]. Available: https://arxiv.org/abs/2406.06578

[3] G. Sousa, D. C. G. Pedronette, J. P. Papa, and I. R. Guilherme, "SMS Spam Detection Through Skip-gram Embeddings and Shallow Networks," in Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 2021, pp. 4193-4201. [Online]. Available: https://aclanthology.org/2021.findings-acl.367/

[4] B. Reaves and P. Traynor, "Detecting SMS Spam in the Age of Legitimate Bulk Messaging," in Proc. 9th ACM Conf. on Security & Privacy in Wireless and Mobile Networks (WiSec), Darmstadt, Germany, 2016, pp. 165-170.

[5] M. Saleh and I. Mahmoud, "Artificial Immune System-Based SMS Spam Detection," in Proc. IEEE Congress on Evolutionary Computation (CEC), Wellington, New Zealand, 2019, pp. 1245-1252.

[6] H. Ulus, A. Ozcan, and B. Yildiz, "Transfer Learning in SMS Spam Detection: Augmentation and Stacking," in Proc. IEEE Global Communications Conf. (GLOBECOM), Abu Dhabi, UAE, 2022, pp. 2310-2315.

[7] T. Silpa and N. Prasad, "Meta-Classifiers for SMS Spam Detection Using Multinomial Naïve Bayes and SVM," Journal of Applied Computing Research, vol. 5, no. 3, pp. 77-85, 2023.

[8] M. Ahmed, L. Yang, and X. Wu, "Machine Learning-Based Spam Detection in IoT Networks," IEEE Internet of Things Journal, vol. 9, no. 8, pp. 7145-7156, 2022.

[9] T. Bari, J. Lee, and M. Kim, "Natural Language Processing-Based Spam Classification for SMS and Email," in Proc. IEEE Conf. on Artificial Intelligence (AI), 2023, pp. 198-204.

[10] A. Uddin, S. Hossain, and M. Rahman, "Transformer-Based SMS Spam Detection with Explainable AI," arXiv preprint arXiv:2405.08026, 2024. [Online]. Available: https://arxiv.org/abs/2405.08026

[11] S. Gadde, R. Singh, and A. Nayak, "Deep Learning Models for SMS Spam Detection: A Comparative Analysis," in Proc. IEEE Int. Conf. on Data Science and Engineering, 2021, pp. 56-62.

[12] A. A. Gedam and R. R. Banchhor, "SMS Spam Detection using Machine Learning Techniques," *2023 IEEE International Conference on Intelligent Systems*, pp. 1–6, 2023, doi: 10.1109/ICIS57433.2023.000

[13] Y. Alghofaili, A. Albattah, and M. A. Rassam, ‘‘A financial fraud detection model based on LSTM deep learning technique,’’ J. Appl. Secur. Res., vol. 15, no. 4, pp. 498–516, Oct. 2020

[14] S. A. Ebiaredoh-Mienye, E. Esenogho, and T. G. Swart, ‘‘Artificial neural network technique for improving prediction of credit card default: A stacked sparse autoencoder approach,’’ Int. J. Electr. Comput. Eng. (IJECE), vol. 11, no. 5, p. 4392, Oct. 2021.

[15] M. Mukherjee and M. Khushi, ‘‘SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features,’’ Appl. Syst. Innov., vol. 4, no. 1, p. 18, Mar. 2021

979-8-3315-9610-1/25/$31.00 ©2025 IEEE