© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1717972

Fraud SMS Spam Detection Using Machine Learning

RITHICK S[1], DR. HARIPRIYA V[2]
[1]Dept. *of Computer Science and Information Technology Jain (Deemed-to-be University) Jayanagar 9th Block Campus, Bengaluru*
[2]*Assistant Professor,Department of Computer Science & IT, Jain (Deemed-to-be) University, Bangalore*

Abstract- ***Fraudulent SMS messages and spam attacks have become a major cybersecurity concern due to the rapid growth of mobile communication and digital services. Traditional filtering systems often fail to identify evolving spam patterns, phishing links, and deceptive text messages used in financial fraud and identity theft. This paper presents a Machine Learning (ML)-based Fraud SMS Spam Detection framework capable of automatically classifying messages as spam or legitimate (ham). A structured review of existing ML and Deep Learning approaches is performed, analysing datasets, preprocessing techniques, feature extraction methods, model architectures, and evaluation metrics. The study proposes a comparative framework using Naive Bayes, Support Vector Machine (SVM), Random Forest, Logistic Regression, and Long Short-Term Memory (LSTM) models. The framework focuses on improving detection accuracy, reducing false positives, and enabling real-time spam filtering. The proposed system aims to support secure mobile communication by providing an intelligent and scalable SMS spam detection mechanism.***

Keywords— Fraud SMS Detection, Spam Classification, Machine Learning, Natural Language Processing, Deep Learning, Naive Bayes, LSTM, Cybersecurity.

I. INTRODUCTION

A. Background of the Study

SMS communication remains one of the most widely used digital communication methods due to its simplicity, low cost, and accessibility. However, the increasing use of mobile services has also led to a rapid rise in spam and fraudulent messages. These messages often contain phishing links, fake offers, financial scams, OTP frauds, and malicious content designed to deceive users.

Traditional rule-based spam filtering systems are unable to effectively detect newly evolving spam patterns because fraudsters continuously modify message structures and vocabulary. Machine Learning (ML) and Natural Language Processing (NLP) techniques provide intelligent approaches capable of learning patterns from large SMS datasets and automatically identifying fraudulent messages.

Recent advancements in ML and Deep Learning have improved text classification performance significantly. Algorithms such as Naive Bayes, Random Forest, Support Vector Machine (SVM), Logistic Regression, and LSTM have shown promising results in spam detection tasks. However, challenges such as imbalanced datasets, multilingual spam messages, real-time filtering, and explainability still remain.

B. Problem Statement

Existing SMS spam detection systems mainly rely on keyword-based filtering techniques that fail to identify modern fraud patterns and context-aware spam messages. Many existing models are trained on limited datasets and are not capable of handling real-time evolving spam attacks.

Additionally, current systems often suffer from high false positive rates, where legitimate messages are incorrectly classified as spam. Most approaches also lack interpretability and adaptability to multilingual or dynamically changing SMS content.

Therefore, there is a need for an intelligent ML-based spam detection framework capable of accurately classifying fraudulent SMS messages while maintaining low false positive rates and supporting real-time deployment.

C. Motivation

The motivation for this work arises from the increasing number of cyber fraud incidents through SMS communication. Fraudulent messages related to banking scams, fake rewards, phishing attacks, and

identity theft continue to affect millions of users worldwide.

Developing an intelligent spam detection system can improve user security, protect sensitive information, and reduce financial losses caused by SMS-based cyberattacks. Machine Learning provides the capability to automatically learn spam patterns and improve detection accuracy over time.

D. Objectives of the Study
The objectives of this research are:

- Analyse SMS datasets and identify spam-related patterns.
- Perform text preprocessing and feature extraction using NLP techniques.
- Develop and compare ML and DL models for SMS spam detection.
-
- Improve spam detection accuracy while reducing false positives.
- Design a real-time fraud SMS filtering framework.
- Evaluate models using standard performance metrics such as Accuracy, Precision, Recall, and F1-Score.

E. Contributions of the Paper
The main contributions of this work are: A structured comparative review of recent Machine Learning and Deep Learning approaches for Fraud SMS Spam Detection. Identification of major research gaps related to real-time spam detection, multilingual spam classification, and false positive reduction. Targeted research objectives mapped to the identified research gaps. A proposed end-to-end intelligent ML-based spam detection framework integrating NLP preprocessing, feature extraction, and real-time fraud SMS classification for secure mobile communication systems.

II. LITERATURE REVIEW

A. Classical Machine Learning Approaches
Traditional ML models such as Naive Bayes, Logistic Regression, and Support Vector Machine (SVM) are widely used for SMS spam classification. Naive Bayes performs effectively on text classification tasks because of its probabilistic learning capability. SVM achieves strong classification accuracy by separating spam and ham messages using optimal hyperplanes.

Random Forest and Decision Tree models are also commonly used due to their ability to handle nonlinear relationships and large feature sets.

B. Deep Learning Approaches
Deep Learning techniques such as Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) models have shown improved performance in text classification.

LSTM models are particularly effective because they capture sequential dependencies in SMS text data and understand contextual information better than traditional ML models.

C. NLP-Based Feature Engineering
Natural Language Processing techniques play an important role in SMS spam detection. Common preprocessing steps include tokenization, stop-word removal, stemming, lemmatization, and TF-IDF vectorization.

Feature extraction techniques help convert text messages into numerical representations suitable for ML models.

D. Real-Time Spam Detection Systems
Recent studies focus on deploying real-time spam filtering systems integrated with mobile applications and cloud-based services. However, many systems still face limitations in handling multilingual spam content and adaptive fraud patterns.

III. III.COMPARATIVE ANALYSIS OF EXISTING METHODS

Table I provides a comparison of the effectiveness and methodology aspects of various machine learning techniques used for building energy consumption forecasting. It shows that ensemble and hybrid architectures achieve superior prediction accuracy, while occupancy and operational features remain

systematically absent from most reviewed frameworks.

TABLE I Comparative Summary of Machine Learning Approaches for Building Energy Consumption Forecasting

Ref	Authors (Year)	Application Domain	ML Models Used	Key Findings
G1 – Dataset		Limited and outdated SMS datasets		Use updated and balanced SMS datasets
G2 – Real-Time Detection		Lack of real-time spam filtering		Develop near real-time detection framework
G3 – Multilingual Spam		Existing models focus only on English		Support multilingual SMS classification
G4 – False Positives		Legitimate messages wrongly classified		Improve precision and reduce false positives
G5 – Explainability		Black-box predictions difficult to interpret		Add explainable AI techniques
G6 – Dynamic Spam Patterns		Fraudsters continuously modify spam formats		Adaptive ML-based learning system
G7 – Scalability		Systems fail on large datasets		Design scalable ML pipeline
G8 – Feature Extraction		Weak text preprocessing methods		Use advanced NLP feature engineering
G9 – Model Comparison		Limited comparative analysis		Benchmark multiple ML and DL models
G10 – Security Integration		Lack of integration with mobile security systems		Build deployable fraud detection framework

IV. PROPOSED METHODOLOGY

A. System Architecture

The proposed framework is a five-module end-to-end pipeline: (1) Data Collection — collects SMS messages from public spam datasets and mobile communication records; (2) Preprocessing and Feature Engineering — handles text cleaning, tokenization, stop-word removal, stemming, and converts SMS text into numerical vectors using NLP techniques; (3) Model Training and Comparison — trains five model categories on an 80:10:10 train-validation-test split; (4) Spam Detection Engine — classifies incoming SMS messages as spam or legitimate in near real-time; (5) Analytics and Decision Output — provides spam prediction reports, fraud alerts, and message classification results for secure mobile communication systems.

B. Feature Engineering

Features are drawn from three categories: (1) Textual Features — word frequency, message length, keyword occurrence, special character count, URL presence, and spam-related vocabulary patterns; (2) NLP Features — TF-IDF vectors, Bag of Words representations, token embeddings, and n-gram analysis; (3) Contextual Features — sender information, message timing patterns, repeated message frequency, and suspicious link indicators. Feature importance is evaluated using information gain and model-based feature ranking techniques to identify the most influential spam indicators.

C. Dataset

The dataset comprises spam and legitimate SMS messages collected from publicly available repositories such as the UCI SMS Spam Collection Dataset and Kaggle datasets. The dataset contains thousands of labelled messages classified into spam and ham categories. Each record includes the SMS text content and the corresponding label. Data cleaning addresses duplicate entries, missing values, unwanted symbols, URLs, and irrelevant text components to improve model training quality.

D. Model Comparison Plan

Five model categories are benchmarked: (a) Baseline — Logistic Regression; (b) Classical ML — Naive Bayes, Support Vector Machine (SVM), Random Forest; (c) Deep Learning — Long Short-Term Memory (LSTM); (d) Hybrid Models — NLP-integrated ML classification frameworks; (e) Interpretable Models — explainable spam classification approaches. An 80:10:10 train-validation-test split is applied to ensure fair and consistent model evaluation. Evaluation metrics include Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.

E. Workflow

The implemented workflow consists of the following interconnected stages: (1) SMS data is collected from

spam detection datasets; (2) Data preprocessing removes unwanted text components and performs normalization; (3) Feature engineering extracts textual and NLP-based features; (4) The dataset is divided into training, validation, and testing subsets; (5) All selected ML and DL models are trained and benchmarked; (6) The best-performing

V. IMPLEMENTATION

The planned implementation will benchmark five machine learning models for Fraud SMS Spam Detection using textual and NLP-based features. The implementation pipeline will consist of five sequential phases: data preprocessing, feature engineering, model training, performance evaluation, and spam prediction generation. This section outlines the detailed plan for each phase.

A. Data Preprocessing Plan

The raw dataset will comprise thousands of SMS messages collected from publicly available spam datasets, each containing SMS text content and corresponding spam or ham labels. Text messages will be converted into a standardized format through lowercase conversion, punctuation removal, stop-word elimination, stemming, and lemmatization. Duplicate entries, unwanted symbols, and irrelevant characters will be removed to improve dataset quality. URLs, numerical patterns, and special characters commonly associated with fraudulent messages will also be extracted and processed as additional indicators for spam detection.

B. Feature Engineering Plan

The preprocessed dataset will be enriched with five categories of engineered features: (i) textual features — word frequency, message length, spam keyword occurrence, URL count, and special character frequency; (ii) NLP representations — Bag of Words (BoW), TF-IDF vectors, and n-gram analysis; (iii) contextual indicators — suspicious sender patterns, repeated message frequency, and phishing-related terms; (iv) semantic features — token embeddings and contextual relationships between words; (v) statistical features — probability distributions of spam-related vocabulary. All extracted features will be normalized using standard scaling techniques to improve model performance and consistency.

C. Training and Testing Strategy

To ensure fair evaluation of SMS spam detection performance, the dataset will be divided into training and testing subsets using an 80:20 split ratio. Cross-validation techniques will be applied to reduce overfitting and improve model reliability. This approach ensures that all models are evaluated on unseen SMS messages while maintaining balanced distributions of spam and legitimate messages. All five models will be trained on an identical feature set and tested on the same held-out partition to enable a fair and consistent comparison.

D. Models to be Implemented

Five machine learning models will be trained and compared on the same feature set and test partition: (1) Logistic Regression — a baseline linear classification model for SMS spam detection; (2) Naive Bayes — a probabilistic classifier widely used in text classification tasks due to its simplicity and efficiency; (3) Support Vector Machine (SVM) — configured with optimized kernel parameters for accurate spam message classification; (4) Random Forest — an ensemble learning model consisting of multiple decision trees to improve classification accuracy and reduce overfitting; (5) Long Short-Term Memory (LSTM) — a Deep Learning sequential model trained to capture contextual relationships and sequential dependencies in SMS text data.

E. Evaluation Metrics

The spam classification performance of all five models will be evaluated using five standard metrics: Accuracy to measure the overall classification correctness; Precision to quantify the percentage of correctly predicted spam messages; Recall to measure the system's ability to identify all fraudulent SMS messages; F1-Score to provide a balanced evaluation of precision and recall; and Confusion Matrix to analyse classification errors and false positive rates. These metrics collectively cover classification accuracy, spam detection capability, and prediction reliability, providing a comprehensive and standardized basis for model comparison in Fraud SMS Spam Detection systems.

VI. EXPECTED RESULTS AND DISCUSSION

The implementation has not yet been executed; the results and discussion presented in this section are therefore anticipated outcomes grounded in the literature review and the theoretical properties of the five selected models. Empirical results will be reported in the final version of this paper upon completion of the implementation.

A. Anticipated Model Performance

Based on evidence from the reviewed literature, the Long Short-Term Memory (LSTM) model is expected to achieve the best overall performance across all evaluation metrics, owing to its ability to capture sequential dependencies and contextual relationships in SMS text data. Random Forest is anticipated to rank second due to its strong ensemble classification capability and robustness against overfitting. Support Vector Machine (SVM) is expected to provide strong classification accuracy for high-dimensional textual features. Naive Bayes is anticipated to perform efficiently with fast prediction speed and reliable spam classification performance in text-based datasets. Logistic Regression, as the baseline model, is expected to exhibit comparatively lower accuracy due to its limited capability in handling complex nonlinear spam patterns.

B. Anticipated Spam Detection Patterns

Fraudulent SMS messages are expected to exhibit identifiable textual and structural patterns such as repeated spam keywords, suspicious URLs, abnormal punctuation usage, promotional phrases, and phishing-related content. Spam messages are also anticipated to contain shorter response-triggering statements and urgent action requests commonly associated with financial scams and fake offers. Legitimate messages are expected to display more natural language patterns with fewer suspicious indicators. These textual differences collectively motivate the inclusion of NLP preprocessing, keyword extraction, and contextual feature engineering techniques in the proposed framework, and are expected to significantly improve spam detection accuracy compared to traditional keyword-based filtering systems.

C. Anticipated Feature Importance

Based on the literature, textual features such as spam keyword occurrence, URL presence, message length, and suspicious phrase frequency are expected to emerge as the strongest predictors across tree-based models. Among contextual indicators, repeated sender patterns and phishing-related vocabulary are anticipated to rank among the top features in Random Forest feature importance analysis, validating the NLP-based feature engineering strategy described in Section IV-B. TF-IDF vectors and n-gram representations are expected to contribute significantly toward distinguishing spam and legitimate messages. Explainable AI techniques will be applied to quantify and visualise feature importance, providing interpretable outputs for spam detection analysis.

D. Anticipated Comparative Insights

The LSTM model is anticipated to outperform all traditional ML models, particularly in detecting complex and dynamically changing spam patterns where contextual understanding is important. However, the performance difference between LSTM and Random Forest is expected to be moderate relative to the substantially higher computational complexity of Deep Learning models. Random Forest combined with NLP-based features is therefore anticipated to represent the strongest interpretable alternative for practical deployment. Logistic Regression's anticipated lower performance is expected to confirm the nonlinear and evolving nature of fraudulent SMS patterns and validate the model selection strategy of the proposed framework.

E. Discussion

The five-model comparative framework is designed to generate evidence-based answers to the research questions identified in Section IV. By benchmarking a Deep Learning model (LSTM), ensemble methods (Random Forest), kernel-based methods (SVM), probabilistic classifiers (Naive Bayes), and a linear baseline (Logistic Regression) on a consistent SMS spam dataset, the implementation will provide a direct comparative analysis for Fraud SMS Spam Detection systems. The anticipated advantage of NLP-integrated models over traditional filtering approaches will directly address challenges related to dynamic spam patterns, false positives, and real-time

detection. Prediction errors are anticipated to occur primarily in highly ambiguous or newly emerging spam messages where historical training patterns provide limited predictive information. Identifying and analysing these high-uncertainty cases will support the development of adaptive and intelligent spam filtering systems in future work.

VII. CONCLUSION

This paper presents a structured review of recent Machine Learning and Deep Learning approaches for Fraud SMS Spam Detection and a planned comparative framework for five ML models targeting fraudulent and spam SMS classification — a domain of increasing importance in modern cybersecurity and mobile communication systems. Major research gaps were identified, with the most significant being limited real-time spam detection capability, high false positive rates, lack of multilingual spam handling, insufficient explainability, and limited adaptability to dynamically evolving fraud patterns. The proposed intelligent ML-based framework, comprising Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest, and Long Short-Term Memory (LSTM), directly addresses these gaps through systematic model comparison, NLP-integrated feature engineering, explainable spam classification, and a near real-time SMS filtering pipeline.

The NLP-based feature engineering strategy — incorporating spam keyword analysis, TF-IDF representations, URL detection, suspicious phrase identification, and contextual text features — is hypothesised to yield measurable improvements in spam detection accuracy over traditional keyword-based filtering systems, with significant reductions in false positive classifications anticipated based on comparable studies. The LSTM model is anticipated to achieve the highest spam classification accuracy, while Random Forest combined with NLP-based feature analysis is expected to provide the strongest interpretable alternative for practical deployment in mobile security systems. Empirical validation of these hypotheses will be conducted in the next phase of this research, with results to be reported in the final version of this paper. Future work will additionally focus on multilingual spam detection, integration with real-time mobile communication platforms, phishing URL analysis, and adaptive AI systems capable of learning newly emerging fraud patterns automatically.

VIII. FUTURE SCOPE

Future implementation will focus on: (1) empirical benchmarking and comparative evaluation of all five model categories on real-world SMS spam datasets containing fraudulent and legitimate message patterns; (2) live spam detection system validation using real-time SMS streams and mobile communication platforms; (3) multilingual SMS dataset collection across different languages and regional communication patterns to improve model generalisability; (4) integration of privacy-preserving and secure AI techniques for protecting sensitive user communication data during spam analysis; (5) development of an end-to-end real-time automated fraud alert system integrated with mobile security and messaging platforms; and (6) deeper exploration of Explainable AI (XAI) techniques to generate interpretable and transparent spam classification results for cybersecurity and mobile communication applications.

REFERENCES.

[1] A. Almeida et al., “SMS Spam Filtering using Machine Learning Techniques,” IEEE, 2023.

[2] S. Gupta et al., “Fraud SMS Detection using Natural Language Processing and Machine Learning,” ICSCC, IEEE, 2024.

[3] R. Sharma and P. Verma, “Performance Evaluation of ML Algorithms for SMS Spam Classification,” MRIE, IEEE, 2025.

[4] K. Reddy et al., “Comparative Analysis of Deep Learning Models for Spam Message Detection,” IEEE, 2026.

[5] A. Thakur et al., “Predictive Analysis of Fraudulent SMS Detection using ML Techniques,” ICSSES, IEEE, 2023.

[6] P. Bhamare et al., “Machine Learning-based SMS Spam Detection System,” ICCUBEA, IEEE, 2024.

[7] T. C. Brito and M. A. Brito, "Spam Message Filtering using Artificial Intelligence Methods," IEEE, 2022.

[8] S. Rajalakshmi et al., "SMS Spam Detection and Mobile Security using Machine Learning," IEEE ICSCSS, 2024.

[9] M. Bhandarkar et al., "Text Classification Techniques for Fraud SMS Detection," ICCUBEA, IEEE, 2023.

[10] M. Venkatesh et al., "Detection of Mobile Spam Messages using ML Algorithms," ICESC, IEEE, 2022.

[11] H. Haque et al., "Spam Detection in Mobile Communication using Machine Learning," IEMTRONICS, IEEE, 2021.

[12] L. Raju et al., "Cost Effective Mobile Spam Detection using IoT and AI," ICEES, IEEE, 2025.

[13] K. Vignesh et al., "SMS Spam Detection using LSTM-based Deep Learning Models," ICECMSN, IEEE, 2025.

[14] R. Mathumitha et al., "SVM-based Classification for Fraud SMS Detection," ICCCNT, IEEE, 2023.

[15] Siranjeevi R, "Hybrid Machine Learning Models for Spam Message Prediction," ICECST, IEEE, 2025.

[16] N. Chauhan et al., "A Comparative Analysis for Short and Long Text Spam Classification," ICCICA, IEEE, 2024.

[17] J. Gaboitaolelwe et al., "Text Message Spam Prediction Using Machine Learning," SmartNets, IEEE, 2022.

[18] M. A. M. Hasan et al., "Enhancing Cyber Fraud Detection through Advanced ML Techniques," OEEC, IEEE, 2025.

[19] R. Halder and K. F. Ahmed, "Multi-Architecture Deep Learning Approach for Enhanced SMS Spam Detection," SEPOC, IEEE, 2025.

[20] D. Syed et al., "Hybrid Deep Learning Models for Mobile Spam Detection," IEEE Access, 2021.

[21] C. Zongo et al., "Machine Learning Approaches for SMS Fraud Detection," IEEE AFRICON, 2023.

[22] A. Jozi et al., "Neuro-Fuzzy Inference Systems for Spam Message Classification," IEEE EEEIC, 2020.

[23] V. Srivastava et al., "Review on Machine Learning Approaches for Spam Detection," ISML, IEEE, 2024.

[24] A. Kidd et al., "Fraud Detection using ML and Text Activity Metrics," CSECS, IEEE, 2025.

[25] R. Yadav et al., "Prediction and Classification of Fraudulent SMS Messages," ICTMIM, IEEE, 2025.