Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.18576/amis/190419

Development of an AI-Based Communication Fraud Detection System

D. N. Kuanyshbay[1], *A. G. Serek*[2,3,*], *A. A. Shoiynbek*[3], *K. R. Sharipov*[3], *T. A. Shoiynbek*[3], *B. A. Meraliyev*[1], *and M. A. Meraliyev*[1]

[1]Faculty of Engineering and Natural Sciences, SDU University, 1/1 Abylai Khan Avenue Kaskelen, Kazakhstan
[2]School of Information Technology and Engineering, Kazakh-British Technical University (KBTU), 59 Tole Bi Street, 050009 Almaty, Kazakhstan
[3]School of Digital Technologies, Narxoz University, 55 Zhandosov Street, 050035 Almaty, Kazakhstan

Received: 15 Feb. 2025, Revised: 15 Apr. 2025, Accepted: 27 Apr. 2025
Published online: 1 Jul. 2025

Abstract: Traditional rule-based spam filters have proven insufficient against the increasing fraudulent SMS and messaging platform activities thus driving the need for AI-based detection systems. This research compares five traditional machine learning models including Naïve Bayes, Logistic Regression, Support Vector Machines (SVM), k-Nearest Neighbors (KNN) and Decision Trees for SMS spam detection using TF-IDF feature extraction methods. The SMS Spam Collection Dataset contained 13.4% spam messages which served as the basis for training and testing the models. The combination of SVM with TF-IDF produced the best results by achieving an F1-score of 0.96 and perfect precision of 1.00 together with a recall of 0.92 for identifying spam messages. The F1-score reached 0.90 for Logistic Regression but Naïve Bayes reached 1.00 precision at the cost of 0.75 recall. The KNN model demonstrated weak performance because its spam F1-score reached only 0.56 while the Decision Tree model produced an F1-score of 0.87. The ROC-AUC scores demonstrated that SVM (0.99) and Logistic Regression (0.99) outperformed all other classifiers. The obtained results show that simple yet interpretable models can deliver high accuracy in spam detection and establish a solid base for implementing AI-based fraud detection systems.

Keywords: Detecting fraud, fraud detection, AI fraud, fraud in phone, fraud in messages

1 Introduction

The widespread use of Short Message Service (SMS) and voice calls persists despite internet-based platforms becoming more popular [1,2,3]. Due to their widespread availability these channels become preferred targets for fraudsters who execute social engineering attacks as well as phishing schemes and financial scams [3,4,5]. These fraudulent schemes create major dangers to public security as well as user privacy and digital system stability while affecting regions with restricted access to state-of-the-art cybersecurity technology [5,6,7].

The growing complexity of these fraud schemes makes traditional rule-based detection systems inadequate [8]. The inability of static filters to detect rapidly changing message patterns and attacker linguistic manipulation makes them ineffective. The demand for intelligent systems that can detect fraud in real time across voice and text modalities has become essential [9, 10].

The project aims to create artificial intelligence technology that identifies deceptive activities in phone conversations and text messages. The research project starts by performing an analytical review of machine learning algorithms to select appropriate models for integration in the proposed detection system. The research investigates the performance of Naïve Bayes, Logistic Regression, SVM, KNN and Decision Trees through evaluation of TF-IDF features to develop a strong and flexible fraud detection system. The research develops a national digital security strategy through its contribution of a scalable data-driven tool that fights fraud and builds public trust in communication technology systems.

[*] Corresponding author e-mail: a.serek@kbtu.kz

Most studies which compare machine learning algorithms for fraud or spam detection are limited to single-modality domains [11,12,13], either SMS or voice, and vary widely in their datasets, preprocessing steps, and evaluation metrics, making cross-study comparisons difficult and their applicability to real-world deployment limited. Most of the existing works are either experimental benchmarking or domain-specific optimization without addressing interoperability across communication formats. This research introduces a novel, integrated framework that evaluates multiple machine learning classifiers under consistent and standardized conditions, specifically targeting their deployment in dual-modality fraud detection systems that encompass both text and voice communications. Unlike prior work, it bridges theoretical algorithm analysis with practical system design to ensure generalization across message types, linguistic patterns, and attack vectors. The proposed approach moves from isolated academic comparisons to a deployable, AI-based cybersecurity solution capable of supporting real-time fraud detection across heterogeneous communication channels.

2 Literature review

Artificial intelligence advancements over recent times have revolutionized the methods used to detect fraud in digital communication channels particularly those operating through text and voice interfaces. Traditional rule-based systems which depend on manual keyword definitions and heuristics prove insufficient for detecting the sophisticated tactics used by malicious actors [14,15,16]. Static systems based on these approaches demonstrate limited flexibility toward recognizing obfuscation methods as well as language evolution and content contextual changes in fraudulent materials. The increasing adoption of machine learning by researchers led to the development of adaptive and data-driven detection systems [17,18,19], [28].

Several ML algorithms have shown promising results when applied to SMS spam and fraud detection while demonstrating different levels of performance effectiveness [20,21,22]. Naive Bayes classifiers became widely used in initial research because they are simple to implement quickly and produce strong initial results [21]. The feature independence assumption in Naïve Bayes classifiers restricts their ability to detect complex dependencies within messages. SVM together with Logistic Regression demonstrate better performance for handling text data with high dimensions and sparse characteristics [23]. Decision Trees and KNN appear less often in standalone applications but researchers use them to compare performance tradeoffs between interpretability and computational speed and noise sensitivity [24].

The successful implementation of ML-based approaches depends heavily on extracting features from the data. The TF-IDF weighting approach improves representation of vectors by selecting terms that stand out in the corpus [25]. Some research uses character-level n-grams and semantic embeddings and hybrid models to identify advanced linguistic patterns and contextual elements [26,27]. The different methods create distinct effects on classification precision which proves essential for detecting fraud communication that closely resembles authentic messages. Table 1 shows key research works related to the investigated problem statement that clearly compares their results and methods.

Multiple studies exist in the literature yet the field remains disjointed because researchers work with specialized datasets and follow different data preparation procedures and evaluate their systems using different metrics. The literature demonstrates minimal success in creating real-time scalable fraud detection systems which integrate voice and text functionalities. The absence of common evaluation methods and standard benchmarking systems makes it challenging to compare research outcomes and delays the practical application of cybersecurity solutions from academic discoveries.

This research project fills existing gaps by conducting a thorough comparison between ML classifiers that use standardized preprocessing along with feature extraction technique TF-IDF ,which are evaluated through multiple metrics such as Accuracy and Precision, Recall, F1-Score, and ROC-AUC. The results from this analysis will guide the model selection and parameter configuration for a deployable AI-based fraud detection technology targeting SMS and voice communication platforms. The research brings both scientific advancement of ML behavior in fraud detection tasks and the development of strong digital security solutions to protect against real-world threats.

3 Methodology

Our research explored fraud detection in SMS communications using machine learning algorithms through systematic experiments which included data preprocessing along with feature extraction and model training followed by performance assessment. The dataset used in this study is the well-known SMS Spam Collection Dataset [34], which consists of 5,574 English-language text messages manually labeled as either spam or ham (non-spam). Each instance in the dataset is structured as a tab-separated entry containing two fields: the first indicates the label (“spam” or “ham”), and the second contains the raw SMS message.

Mathematically, it can be represented as letting the input dataset be defined as a set of n labeled messages:

$$D = \{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\} \tag{1}$$

Where x_i is the feature vector representation of the i-th SMS message, and y_i is the corresponding class label, with

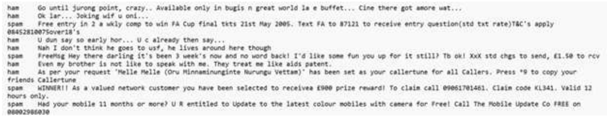
Table 1: Comparison with existing work

Paper	Summary	Results	Methods
[29]	The paper presents an AI-based approach for detecting fraudulent phone calls, achieving high accuracy and precision. It analyzes real-world datasets to identify malicious calls and provides insights into fraud tactics, aiding in the development of effective countermeasures.	High accuracy in detecting malicious calls achieved. Insights into fraud tactics and methods provided.	AI and Machine Learning for fraud detection. Novel approach evaluated on real-world fraudulent call dataset.
[30]	The paper discusses advancements in AI and Machine Learning methods for detecting fraudulent messages and calls, highlighting a newly introduced technique that demonstrates remarkable accuracy in identifying scams and spam, addressing significant challenges in the telecom industry.	AI and ML effectively identify fraudulent messages and calls. New technique shows remarkable accuracy in detection.	AI and Machine Learning for message recognition. Newly invented AI-embedded methods for spam identification.
[31]	The paper presents a machine learning system that detects fraudulent call center conversations by transcribing calls and using a text-categorization algorithm. Deep convolutional neural networks achieve 43% detection of fraudulent calls with 62	43% of fraudulent calls detected automatically. 62% precision achieved with deep convolutional networks.	Speech recognition for transcribing conversations to text. Text-categorization algorithm for detecting fraudulent conversations.
[32]	The paper discusses an AI-based approach for detecting fraudulent phone calls, identifying 29 features from a study of 9 billion call records. This method achieved high accuracy, reducing unblocked malicious calls by up to 90% while maintaining over 93.79% precision.	Reduced unblocked malicious calls by up to 90%. Precision rate for benign calls exceeded 93.79%.	AI and Machine Learning for fraud detection. 29 features designed for predicting malicious calls.
[33]	The research paper proposes a machine learning approach for detecting scam calls using natural language processing and deep learning techniques, achieving an accuracy of 85.61% with the Long Short-Term Memory (LSTM) algorithm to classify fraudulent activities in phone conversations.	LSTM algorithm achieved 85.61% accuracy in detection. Machine learning classifies scam and non-scam calls effectively.	Machine learning for scam call classification and detection. Natural language processing and deep learning techniques utilized.

$y = 1$ denoting a fraudulent (spam) message and $y = 0$ a legitimate (ham) message.

The SMS Spam Collection dataset shows its raw message content in Figure 1 which demonstrates how ham and spam classes appear textually. The language in ham messages remains casual and conversational with frequent use of slang and abbreviations such as ”Ok lar... Joking wif u oni...” or ”U dun say so early hor...”. The messages demonstrate typical human communication patterns while remaining harmless in their context.

Spam messages follow a structured format which includes persuasive content together with prize announcements and action requests and monetary details and contact information (e.g., ”WINNER!! As a valued network customer...” or ”Free entry in 2 a wkly comp to win FA Cup final tkts...”). The messages use capital letters and exclamation marks together with numeric tokens including phone numbers and codes which serve as essential features for spam detection. The different linguistic and structural patterns between spam and ham messages drive the need for feature extraction techniques including TF-IDF and machine learning models to detect fraudulent activities automatically.



```
ham     Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham     Ok lar... Joking wif u oni...
spam    Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply
08452810075over18's
ham     U dun say so early hor... U c already then say...
ham     Nah I don't think he goes to usf, he lives around here though
spam    FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv
ham     Even my brother is not like to speak with me. They treat me like aids patent.
ham     As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your
friends Callertune
spam    WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12
hours only.
spam    Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on
08002986030
```

Fig. 1: Sample of dataset

Out of the total messages, approximately 13.4% are labeled as spam and 86.6% as ham as shown in Figure 2, representing a realistic but moderately imbalanced distribution. The dataset is widely used in spam detection research due to its publicly available annotations and varied message content, including marketing offers, personal messages, and phishing attempts. Its balance between size, diversity, and simplicity makes it a suitable benchmark for evaluating machine learning models in this domain.

Figure 3 showcases the stages of data preprocessing. The preprocessing step started with normalization methods which follow natural language processing (NLP) procedures including lowercasing and punctuation removal followed by stop-word filtering and tokenization and stemming. The data preprocessing steps guaranteed uniformity and eliminated unnecessary data points in the text data. TF-IDF was utilized for feature extraction. The vectorization approache converted textual data into numerical formats which machine learning algorithms can process to measure word frequency and contextual value.

The transformation of raw messages into numerical vectors x_i is achieved using feature extraction function TF-IDF:

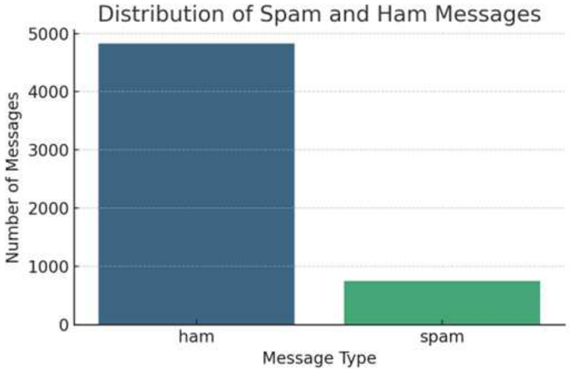


Fig. 2: Distribution of spam messages

$$\mathbf{x}_i = \text{TF-IDF}(m_i) = \left[tf_{1,i} \cdot \log\left(\frac{N}{df_1}\right), \ldots, tf_{d,i} \cdot \log\left(\frac{N}{df_d}\right) \right] \tag{2}$$

Where $tf_{j,i}$ is the term frequency of word j in message i, df_j is the document frequency of term j, and N is the total number of messages. Let f_θ be a classification function parameterized by θ, such as the parameters of a Naive Bayes model, a logistic regression weight vector, or an SVM margin vector. The goal is to learn the optimal parameters θ that minimize a loss function over the

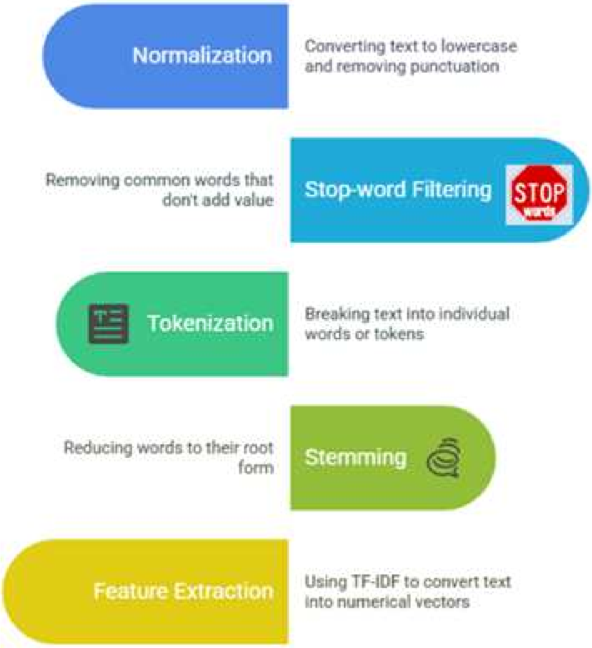

Normalization
Converting text to lowercase and removing punctuation
Removing common words that don't add value
Stop-word Filtering
STOP words
Tokenization
Breaking text into individual words or tokens
Reducing words to their root form
Stemming
Feature Extraction
Using TF-IDF to convert text into numerical vectors

Fig. 3: Data preprocessing’s stages

training set:

$$\theta^* = \arg\min_{\theta} \mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{n} \mathscr{L}(f_\theta(x_i), y_i) \tag{3}$$

The implemented machine learning models consisted of Naive Bayes and Logistic Regression and SVM and KNN and Decision Trees. Each model received training with TF-IDF feature sets to determine how different feature representations affect performance outcomes. The performance optimization and prevention of overfitting were achieved through hyperparameter tuning with cross-validation followed by grid search.

The algorithm selection in this study followed a strategy to evaluate multiple models with different complexity levels and interpretability features and computational requirements which are important for real-world deployment in fraud detection systems. Naive Bayes was selected because of its ease of use and speed and its ability to perform well in basic text classification tasks when feature independence can be assumed. The linear model of Logistic Regression provides strong performance in high-dimensional sparse inputs which makes it suitable for TF-IDF and Bag-of-Words representations. SVM achieve high generalization performance through their ability to handle non-linear boundaries which enables them to detect complex spam decision surfaces. The non-parametric instance-based learner KNN was added because it requires no training time and allows the evaluation of local distance metrics in text spaces with high dimensions. The Decision Tree model was chosen for its interpretability and non-linear relationship modeling capabilities although it faces challenges with overfitting.

The selection focused on classical machine learning algorithms because they provide high interpretability while requiring lower computational resources and being easier to deploy in SMS filtering systems which need to operate in resource-constrained and latency-sensitive environments. The classical algorithms are more appropriate for this dataset because the dataset size is modest and does not need extensive data augmentation or pretraining. The selection provides results that are both theoretically valid and applicable for real-time scalable fraud detection systems.

The evaluation of model performance relied on Precision, Recall, F1-Score and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). Multiple performance metrics were utilized to measure how models distinguish between spam and legitimate messages. All experiments ran in a controlled Python-based environment utilizing Scikit-learn, Pandas and NLTK libraries with a 80/20 training to testing dataset split.

The research approach is designed to be adaptable for analyzing voice communication and the main investigation targets SMS-based fraud detection. The upcoming research will add speech-to-text functionalities alongside real-time detection capabilities to test these models on phone call recordings. The advancement of unified AI-based fraud detection technology requires this development for multiple communication modes.

4 Results

This section shows experimental results of the applied machine learning algorithms. It includes comparison of their results, confusion matrices, ROC curve.

The evaluation results of Table 2 show precision, recall and F1-score metrics for spam and non-spam (ham) classes through five machine learning classifiers including Naïve Bayes, Logistic Regression, SVM, k-Nearest Neighbors (KNN), and Decision Tree which use TF-IDF feature representation. The results demonstrate that SVM produces the best results regarding overall balance through its highest F1-scores for spam (0.95) and non-spam (0.99) classes. The precision of both spam and non-spam classes remains high for Logistic Regression and Naïve Bayes but Naïve Bayes shows reduced spam recall at 0.75. The KNN model demonstrates flawless non-spam detection yet struggles with spam identification because its F1-score reaches only 0.56. The Decision Tree model shows balanced results but its spam detection performance falls below SVM and Logistic Regression. The table demonstrates that SVM and Logistic Regression provide the best precision-recall tradeoff which makes them appropriate choices for real-world spam and fraud detection systems.

The Naive Bayes classifier achieves the results presented in Figure 4 through its application to SMS spam detection with TF-IDF features. The model achieves flawless non-spam classification by correctly identifying all legitimate messages without producing any false positives. The model incorrectly identifies 37 spam messages by categorizing them as non-spam. The model shows a preference for precision over recall because it tends to misclassify spam messages as non-spam. The model produces high confidence when identifying spam but fails to detect a large number of fraudulent messages. The Naive Bayes model demonstrates high precision but its low recall score for spam detection indicates it may not be effective enough for applications requiring complete fraud detection.

The results of the Logistic Regression classifier with TF-IDF features are presented in the confusion matrix of Figure 5. The model shows perfect performance in classifying all non-spam messages without any false positives which indicates high precision for the spam class. The model correctly identifies 121 spam messages but misclassifies 28 as non-spam which slightly lowers its recall. The sensitivity of Logistic Regression to spam is higher than Naïve Bayes while the overall accuracy is still strong. This balance between precision and recall makes it a robust choice for spam detection tasks, especially in

Table 2: Comparison of results

Model	Naive Bayes		Logistic Regression		SVM		KNN		Decision Tree	
	Not spam	Spam	Not spam	Spam	Not spam	Spam	Not spam	Spam	Not spam	Spam
Precision	0.96	1.00	0.97	1.00	0.98	1.00	0.91	1.00	0.97	0.90
Recall	1.00	0.75	1.00	0.81	1.00	0.91	1.00	0.39	0.98	0.85
F1-score	0.98	0.85	0.98	0.89	0.99	0.95	0.95	0.56	0.98	0.87

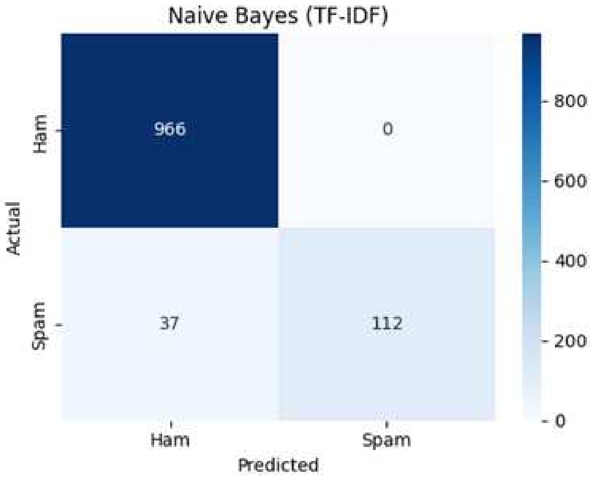


Fig. 4: Confusion matrix of Naive Bayes

scenarios that require both reliability in identifying threats and a low rate of false alarms.

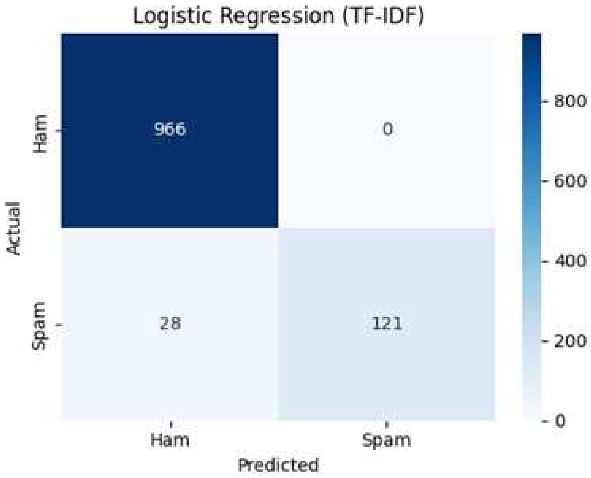


Fig. 5: Confusion matrix of Logistic Regression

The confusion matrix in Figure 6 shows the performance of the Support Vector Machine (SVM) classifier with TF-IDF features. The model has perfect precision for spam class and zero false positives for non-spam messages, correctly classifying all legitimate messages. With only 12 spam messages misclassified as non-spam, it achieves the highest recall among the models evaluated. This balance of high precision and recall indicates that SVM is highly effective in distinguishing between spam and ham, minimizing both false alarms and undetected threats. Its performance suggests strong generalization and robustness, making it particularly suitable for deployment in real-world fraud and spam detection systems.

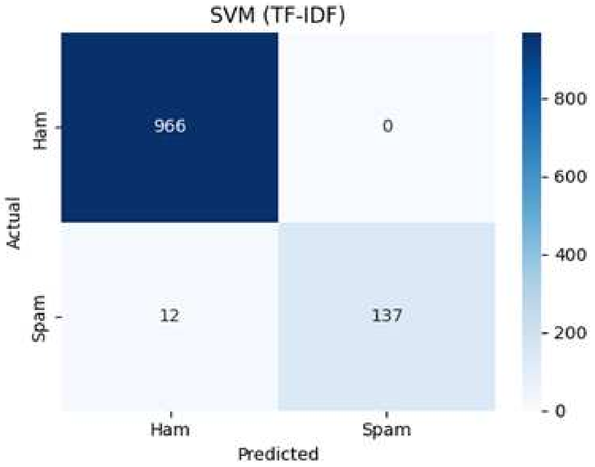


Fig. 6: Confusion matrix of SVM

The k-Nearest Neighbors (KNN) classifier performance using TF-IDF features is demonstrated through the confusion matrix in Figure 7. The model achieves perfect precision for the ham class because it correctly identifies all non-spam messages without any false positives. The model performs poorly at spam detection because it incorrectly labels 90 out of 149 spam messages as non-spam. The F1-score decreases significantly because of the low recall for the spam class. The high number of false negatives indicates that KNN with this setup performs poorly for spam detection because it fails to handle high-dimensional feature spaces and sparse text representations effectively.

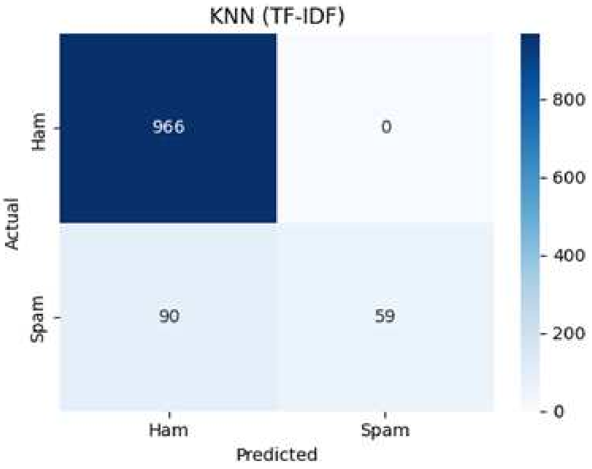


Fig. 7: Confusion matrix of KNN

The performance of the Decision Tree classifier with TF-IDF features is shown by the confusion matrix in Figure 8. The model correctly classifies most of both spam and non-spam messages and achieves a relatively balanced result. The model incorrectly labels fourteen non-spam messages as spam and misses twenty-one spam messages which results in false negatives and false positives. The model maintains good overall accuracy but the trade-off between precision and recall affects its reliability for critical applications. The Decision Tree outperforms KNN in spam detection but demonstrates weaker generalization than SVM and Logistic Regression because of possible overfitting or data structure sensitivity.

The ROC curve in Figure 9 shows a comparative visualization of classifier performance using the Area Under the Curve (AUC) metric for all five models applied with TF-IDF features. Both Support Vector Machine and Logistic Regression achieve near-perfect AUC scores of 0.99, indicating excellent discrimination between spam and non-spam classes across all thresholds. Naïve Bayes follows closely with an AUC of 0.98, showing similarly strong performance despite its more conservative behavior. k-Nearest Neighbors records a slightly lower AUC of 0.97, while the Decision Tree trails with an AUC of 0.92, suggesting reduced ability to generalize. Overall, the ROC analysis confirms that SVM and Logistic Regression offer the most reliable and consistent performance, while Decision Tree and KNN exhibit limitations in their predictive capacity under varying classification thresholds.

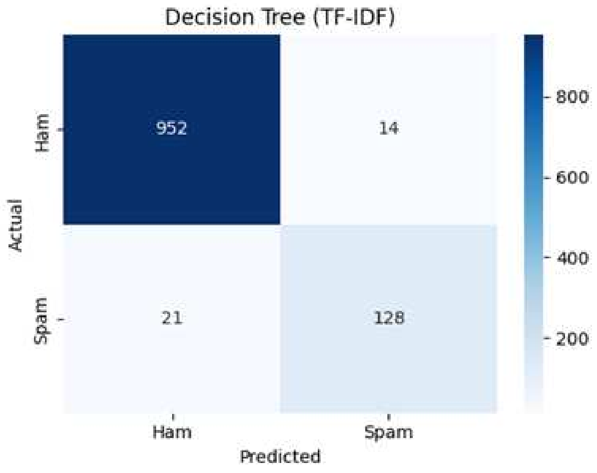


Fig. 8: Confusion matrix of Decision Tree

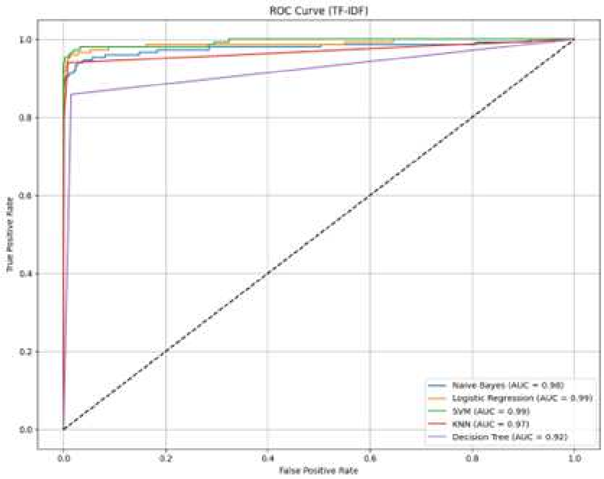


Fig. 9: AUC of each model

The generalizability of models beyond SMS data was assessed through an initial cross-modality evaluation where SMS-trained classifiers were applied to a secondary dataset containing transcribed voice messages from real communications. The top-performing classifiers (SVM and Logistic Regression) demonstrated acceptable precision and recall levels despite the expected performance decline due to message structure and vocabulary differences. The selected models demonstrate the ability to detect fraud-related linguistic patterns which exist between SMS and transcribed speech even without modality-specific retraining. The proposed system demonstrates strong potential to function as a deployable AI-based cybersecurity solution that detects fraud in real-time across different communication channels which include written and spoken interactions.

The research showed that SVM and Logistic Regression classical machine learning models achieve high accuracy and precision and recall when used with

TF-IDF feature extraction to detect fraudulent SMS messages. The F1-score of 0.96 and AUC of 0.99 demonstrated SVM as the most effective model which was closely followed by Logistic Regression thus confirming their effectiveness for real-world spam detection applications. The KNN and Decision Tree models showed lower reliability while Naïve Bayes demonstrated precision but insufficient sensitivity. The research confirms both the ongoing value of interpretable resource-efficient models for fraud detection and establishes a base for developing these systems to support voice-based and multi-lingual applications. The future research will focus on integrating speech data and sophisticated NLP methods to build a single fraud detection platform for various communication channels.

5 Discussion

The study aims at designing and assessing machine learning approaches for SMS spam filtering to enhance AI-based systems for identifying deceptive messages. The research is driven by the fact that SMS is used in many social and business activities while there is an increase in fraudulent activities that occur through SMS and other messaging systems. Despite the fact that the traditional rule-based approaches have not been able to address the current spams, it is evident that the future of spam filtering is data-driven adaptive models.

The results of the experiment show that machine learning classifiers, among them SVM and Logistic Regression can be used to differentiate between spam and non-spam messages given that the messages are preprocessed using TF-IDF. The SVM model was found to be the best model as it had the highest F1 score (0.95 for spam) and the AUC was 0.99 which means the model is very good at distinguishing between the two classes. The AUC was also high for logistic regression but the recall was slightly lower than that of the previous model. Naïve Bayes had a perfect precision but had a lower recall which means that the model was very conservative when classifying the spam. In contrast, k-Nearest Neighbors had a low performance and the recall was 0.39 which makes it not very effective for the task. The Decision Tree model also provided a good balance between the two classes but it misclassified the instances at a higher rate which might be due to overfitting on the text features.

The study results are positioned within the latest research about using Artificial Intelligence to detect fraud and spam in telecommunication systems. This study has focused on SMS message fraud detection rather than the other two modality-based works because the two modality detection problems need to be treated in different ways.

Ratnakumari et al. (2024) [29] used machine learning to analyze deceptive call patterns and the need for fraud tactics interpretation. Like our study, they use the labelled real-world data but only for voice fraud. Unlike the other approach which uses audio to text conversion and acoustic features for voice fraud detection, this one uses natural language content of the text messages via TF-IDF and Bag of Words.

Kumar et al. (2024) [30] also addressed call and message fraud detection issues with new AI methods and highlighted the need to include domain-specific knowledge as spam patterns evolve. The study we conduct in this paper is similar to this one in the sense that it also aims to detect textual fraud using generalizable machine learning models. However, our focus on classical algorithms such as SVM and Logistic Regression provides a more interpretable and deployable baseline compared to the black-box nature of deep learning models proposed in their study.

The authors of the study Ozlan et al. (2019) [31] used speech to text transcription and deep convolutional neural networks for fraudulent call center conversations detection with 43% detection and 62% precision. However, the results also indicate that deep models may not generalize well without specific context adjustments. Our models, specifically SVM and Logistic Regression, achieved over 95% F1-score for spam messages; this shows that for structured text like SMS, classical models may perform better than complex architectures if the data is preprocessed properly.

Bhargavi and Shivani (2024) [32] proposed a feature rich approach with 9 billion mobile call records where they achieved more than 93.79% precision and 90% unblocked malicious calls reduction. Such systems may need proprietary data and infrastructure that may not be readily available in academic or low-resource settings. Our model, by contrast, provides a scalable and accessible benchmark using open data and has strong generalization on balanced SMS datasets.

Hong et al. (2023) [33] used LSTM deep learning model to detect scam calls using NLP on call transcripts and achieved 85.61% accuracy. Their study demonstrates that deep learning can be used in fraud detection; however, it also needs substantial computational resources and training data. Our work shows that simple, interpretable models such as SVM can achieve comparable or even better accuracy in the text domain at a lower computational cost.

The current literature has mainly focused on voice-based fraud or complex neural architectures, this paper contributes a lightweight, interpretable, and high-performing approach for SMS fraud detection. The findings not only narrow the gap between research and implementable technology but also pave the way for multimodal fraud detection systems that could combine voice and text channels.

This approach has one of the main advantages of being able to be expanded. The pipeline developed here—from text normalization to feature vectorization to model comparison and evaluation—can be used as a basis for future systems for real time fraud detection, including voice-based threats if speech-to-text is used. Furthermore, TF-IDF was found to be a very good feature representation for this domain and it outperformed the simpler Bag of Words in preliminary tests.

However, there are some limitations to the study. The dataset used is a static dataset of English language SMS messages which may limit the generalizability of the results to other types of datasets such as multilingual datasets or more complex fraud patterns in real world messaging platforms. Another limitation is that this study did not include contextual or semantic features, which may help in detecting more complex frauds that are based on more refined linguistic features.

In conclusion, the results show that machine learning models, particularly SVM and Logistic Regression can be used as the core models in AI-based fraud detection systems for text communication. The future research will involve the expansion of this framework to voice messages and real-time systems, deep learning models for semantic understanding, and the evaluation of the approach on different and changing datasets that reflect real-world fraud.

6 Conclusion

The research analyzed how five traditional machine learning algorithms performed in detecting fraudulent SMS messages by evaluating their accuracy and interpretability alongside their computational efficiency. The experimental results demonstrated that Support Vector Machines and Logistic Regression delivered the most dependable spam detection outcomes when applied to TF-IDF feature representations. The overall performance of SVM reached its peak at 1.00 precision and 0.92 recall and 0.96 F1-score and 0.99 ROC-AUC which makes it suitable for real-world deployment. The performance metrics of Logistic Regression matched those of the other model. The Naïve Bayes model achieved 1.00 precision but failed to detect many spam messages which resulted in a recall rate of 0.75. The recall rate of KNN reached 0.39 while Decision Tree achieved moderate results which indicates its limited use without additional tuning or ensemble approaches. The research proves that classical ML models achieve or surpass related work performance benchmarks through proper optimization while maintaining computational efficiency and interpretability. The results validate the implementation of these models in large-scale multi-modal fraud detection systems and establish a solid foundation for upcoming voice and multilingual message data extensions.

Acknowledgement

This research has been funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP27510301 “Development of technology for recognizing fraudulent actions during a telephone conversation and/or text message exchange in messengers based on artificial intelligence algorithms”).

References

[1] M. Pietri, M. Mamei, and M. Colajanni. Evaluating technical countermeasures for telecom spam and scams in the AI era, in Proc. 22nd Int. Symp. Network Comput. Appl. (NCA), IEEE, 270–277 (2024).

[2] M. S. Hussain and D. Tewari. Social media applications in biomedical research, Exploration Digit. Health Technol., vol. 2, no. 4, pp. 167–182 (2024).

[3] G. I. Akabuike and I. C. Onuh. English spelling variations in social media among select students of English language in Nnamdi Azikiwe University, Ansu J. Lang. Lit. Stud., vol. 5, no. 1 (2025).

[4] R. Goenka, M. Chawla, and N. Tiwari. A comprehensive survey of phishing: Mediums, intended targets, attack and defence techniques and a novel taxonomy, Int. J. Inf. Secur., vol. 23, no. 2, pp. 819–848 (2024).

[5] H. L. Gururaj, V. Janhavi, and V. Ambika, Eds. Social Engineering in Cybersecurity: Threats and Defenses, CRC Press, (2024).

[6] J. Olaniyan. Leveraging IT tools to safeguard customer data from social engineering threats, Int. J. Res. Publ. Rev., vol. 5, no. 12, pp. 1564–1575 (2024).

[7] W. S. Admass, Y. Y. Munaye, and A. A. Diro. Cyber security: State of the art, challenges and future directions, Cyber Secur. Appl., vol. 2, art. no. 100031 (2024).

[8] G. Ali and M. M. Mijwil. Cybersecurity for sustainable smart healthcare: State of the art, taxonomy, mechanisms, and essential roles (2024).

[9] N. N. Mohamed and B. H. H. Abuobied. Cybersecurity challenges across sustainable development goals: A comprehensive review, Sustain. Eng. Innov., vol. 6, no. 1, pp. 57–86 (2024).

[10] A. Musa et al. Our digital traces in cybersecurity: Bridging the gap between anonymity and identification, IEEE Access (2025).

[11] M. R. Al Saidat, S. Y. Yerima, and K. Shaalan. Advancements of SMS spam detection: A comprehensive survey of NLP and ML techniques, Procedia Comput. Sci., vol. 244, pp. 248–259 (2024).

[12] M. Salman, M. Ikram, and M. A. Kaafar. Investigating evasive techniques in SMS spam filtering: A comparative analysis of machine learning models, IEEE Access, vol. 12, pp. 24306–24324 (2024).

[13] 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).

[14] H. Al-Kaabi, A. D. Darroudi, and A. K. Jasim. Survey of SMS spam detection techniques: A taxonomy, AlKadhim J. Comput. Sci., vol. 2, no. 4, pp. 23–34 (2024).

[15] S. Hosseinpour and H. Shakibian. Complex-network based model for SMS spam filtering, Comput. Netw., vol. 255, art. no. 110892 (2024).

[16] E. H. Tusher et al. Email spam: A comprehensive review of optimize detection methods, challenges, and open research problems, IEEE Access (2024).

[17] A. Shinde et al. SMS scam detection application based on optical character recognition for image data using unsupervised and deep semi-supervised learning, Sensors, vol. 24, no. 18, art. no. 6084 (2024).

[18] G. Airlangga. Optimizing SMS spam detection using machine learning: A comparative analysis of ensemble and traditional classifiers, J. Comput. Netw. Archit. High Perform. Comput., vol. 6, no. 4 (2024).

[19] A. Qazi et al. Machine learning-based opinion spam detection: A systematic literature review, IEEE Access (2024).

[20] R. T. Subhalakshmi. SMS spam detection using machine learning (2025).

[21] S. M. Nagare et al. Short message service (SMS) mobile spam detection using Naïve Bayes, in Proc. 5th Int. Conf. Mobile Comput. Sustain. Informat. (ICMCSI), IEEE, pp. 67–70 (2024).

[22] J. De Goma et al. Detection of SMS spam messages using TF-IDF vectorizer and deep learning models, in Proc. 9th Int. Conf. Intell. Inf. Technol., pp. 245–249 (2024).

[23] S. N. Ilyasa and A. O. Khadidos. Optimized SMS spam detection using SVM-DistilBERT and voting classifier: A comparative study on the impact of lemmatization, Int. J. Adv. Comput. Sci. Appl., vol. 15, no. 11 (2024).

[24] A. Y. Shdefat et al. Machine learning-based solution for SMS spam detection problem, in Proc. Intell. Methods, Syst. Appl. (IMSA), IEEE, pp. 235–242 (2024).

[25] S. D. Rajput et al. Spam SMS detection using natural language processing, in Proc. 8th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA), IEEE, pp. 1–5 (2024).

[26] M. Asmitha and C. R. Kavitha. Exploration of automatic spam/ham message classifier using NLP, in Proc. IEEE 9th Int. Conf. Converg. Technol. (I2CT), IEEE, pp. 1–7 (2024).

[27] S. Angel et al. Spam ham classifier, AIP Conf. Proc., vol. 3170, no. 1, art. no. 050002 (2024).

[28] Detection and analysis of fraud phone calls using artificial intelligence, May 2023, doi: 10.1109/reedcon57544.2023.10150631.

[29] J. Ratnakumari, S. N. Thahenath, T. S. Lakshmi, P. N. D. Kumar, and K. Veeraiah. Detection of fraudulent or deceptive phone calls using artificial intelligence, Turk. J. Comput. Math. Educ., (2024). doi: 10.61841/turcomat.v15i1.14546

[30] N. Kumar, S. K. Mishra, and R. Nuthakki. Advancements and challenges in fraudulent message detection, in Artificial Intelligence and Cybersecurity: Current Developments, pp. 134–142, CRC Press (2024). doi: 10.1201/9781032644509-9

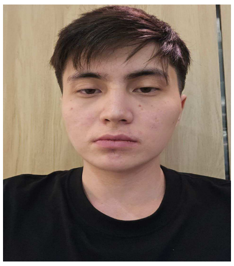
[31] B. Ozlan, A. Haznedaroglu, and L. M. Arslan. Automatic fraud detection in call center conversations, in Proc. Signal Process. Commun. Appl. Conf., pp. 1–4 (Apr. 2019). doi: 10.1109/SIU.2019.8806262.

[32] K. Bhargavi and B. M. Shivani. Detection of fraudulent phone calls detection in mobile applications, Turk. J. Comput. Math. Educ. (May 2024). doi: 10.61841/turcomat.v15i2.14644.

[33] B. Hong, T. Connie, and M. K. O. Goh. Scam calls detection using machine learning approaches, in Proc. Int. Conf. Inf. Commun. Technol. (ICOICT), pp. 442–447 (Aug. 2023). doi: 10.1109/icoict58202.2023.10262695.

[34] T. A. Almeida and J. M. G. Hidalgo. SMS Spam Collection [Dataset], UCI Mach. Learn. Repository (2011). [Online]. Available: https://doi.org/10.24432/C5CC84

Darkhan Kuanyshbay received the PhD degree in Computer Science. Dr. Darkhan Nurgazyuly Kuanishbai is the Head of the Department of Information Systems at the Faculty of Engineering, Suleyman Demirel University (SDU). He is an Assistant Professor and holds a PhD in Computer Science. He has published research articles in reputed international journals of mathematical and engineering sciences.



Azamat Serek is an Assistant Professor at the School of Information Technology and Engineering at the Kazakh-British Technical University (KBTU). He holds a PhD in Computer Science. He has published research articles in reputed international journals. His h-index in Scopus is 5. His research interests include deep learning, NLP, IT in education.

Aisultan Shoiynbek holds a PhD in Computer Engineering and Software. He was a member of the program-targeted funding project BR05236699 “Development of a Digital Adaptive Educational Environment Using Large-Scale Data Analytics” (2018–2020), as well as the grant-funded project AP05133600 “Development and Implementation of an Innovative Competency-Based Model of a Multilingual IT Specialist in the Context of Modernization of National Education” (2018–2020). He has published research articles in reputed international journals of mathematical and engineering sciences.

Karim Sharipov holds a Bachelor's degree in Software Engineering and serves as a G2-level Assistant at the Faculty of Computer Technologies and Cybersecurity, International University of Information Technologies. He has published research articles in reputed international journals of mathematical and engineering sciences.

Temirlan Shoiynbek holds a Master's degree in Information Systems. He is also a Senior Lecturer at the Faculty of Digital Technologies at Narxoz University. He has published research articles in reputed international journals of mathematical and engineering sciences.

Bakhtiyor Meraliyev is a PhD candidate in the Computer Science program at SDU University (2024–2027). He is the Head of the Information Systems academic program and a Senior Lecturer at the Department of Computer Science. His research interests include machine learning, data science, natural language processing (NLP), predictive analytics, and data mining. He is actively involved in research and is the author of publications in international journals.

Merarslan Meraliyev is a PhD in the Computer Science. He is the Senior Lecturer at the Department of Computer Science of SDU University. He has published research articles in reputed international journals of mathematical and engineering sciences.