Identifying Fraudulent Activity in Machine Learning for Telecommunications Employing in CDR

R. Sivaramakrishnan[1], Saravanan Murugesan[2]
[1]Department of Computer Science and Engineering, KPR Institute of Engineering and Technology,
Coimbatore, Tamil Nadu 641407, India
[2]Department of Artificial Intelligence and Data Science, KPR Institute of Engineering and Technology,
Coimbatore, Tamil Nadu 641407, India

Article Info

Article history:

Received Aug 19, 2022
Revised Oct 20, 2022
Accepted Nov 25, 2022

Keywords:

Call detail records
Machine learning
Support vector machine
Finite state machine
Network learning
Components

ABSTRACT

Telecommunication scams are a global issue that costs many customers and service provider businesses a significant amount each year. We provide an efficient and suitable scam user identification technique that depends on the customer's Call Detail Record (CDR) for quick and affordable identification of telecommunication scam clients. The techniques for classification employed in network learning as well as in several scientific and technical domains, such as computer vision, understanding speech, email virus detection, etc., served as the inspiration for the studies we did. The ML and pattern recognition modules make up the two halves of the suggested approach. The ML module uses summary features to identify people using the Support Vector Machine (SVM) technique, which is based on supervised learning. The model identification unit analyzes suspect individuals produced by the machine's instructor using a Finite State Machine (FSM) according to scam user activity. Unauthorized users can be identified after using both components. We implement our approach and evaluate it with actual data. The trials show that the proposed method can achieve exact detection and outperforms the state-of-the-art techniques.

Corresponding Author:

R. Sivaramakrishnan,
Department of Computer Science and Engineering,
KPR Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India.
Email: sivaraamakrishnan2010@gmail.com

1. INTRODUCTION

Research on identifying telecommunication scam users has been conducted in the past few years, and several methods for detecting them have been put out. Fraud user detection often follows several investigation avenues [1]. Through the examination of spoken word or short message service material, the first research line, known as content-based analysis, attempts to identify fraudulent users. Activity-based examination, which is a second research line that focuses on subscriber behavior, looks at how fraud users interact with one another in an attempt to identify fraudulent subscribers. Employing content assessment techniques, four approaches were used to identify telecom scams by taking advantage of the SMS's distinctive terms. A more potent and flexible spam filtering system utilizing SVM and a dictionary was suggested. Presented the neural network approach for detecting fraudulent SMS. To identify fake SMS, techniques for clustering have been widely used. Suggested clustering data utilizing the rocking technique before choosing an approach to construct a classifier.

Especially contrasted with domestic services, increased calling rates encourage fraudsters to end calling abroad on any local provider [2]. Identity Box scam is the practice of re-initiating the global operator's data as a local call and then using unlawful methods to cut off the desired recipient. Due to telecom providers' obligations to protect subscriber privacy, there is only a small amount of information available for exploratory research, which results in a limited number of potential solutions. Typically, fraud administration systems are limited to detecting particular sorts of fraud and are not very good at spotting novel hazards.

Furthermore, the vast collection that the operators keep makes it even more difficult to make decisions about customer segmentation in real time. Furthermore, there is no past information accessible for study because Mobile Box fake subscribers frequently switch SIMs and alter their activity patterns. Only following an exhaustive investigation can a SIM Box fake user be classified as an actual international workshop on Information Systems member.

It is feasible to detect fake subscribers to landline telephone services by examining their setup and usage locations [3]. However, it takes an enormous amount of time and cash to examine every home client in a company such as the Theory of Complexity which boasts thousands of clients at home, to identify all fake customers. Therefore, it is quite demanding to reduce the quantity of consumers to be verified. Figure 1 show that the purpose of this study is to identify trends in the behavior of home and business subscribers by analyzing their CDR and bill data. This will help distinguish between residential subscriptions, which behave similarly to fraudulent clients. Our goal is to identify the actual subscription type as accurately as possible. The loss from a significant portion of telecom revenue can be avoided by identifying membership scams.

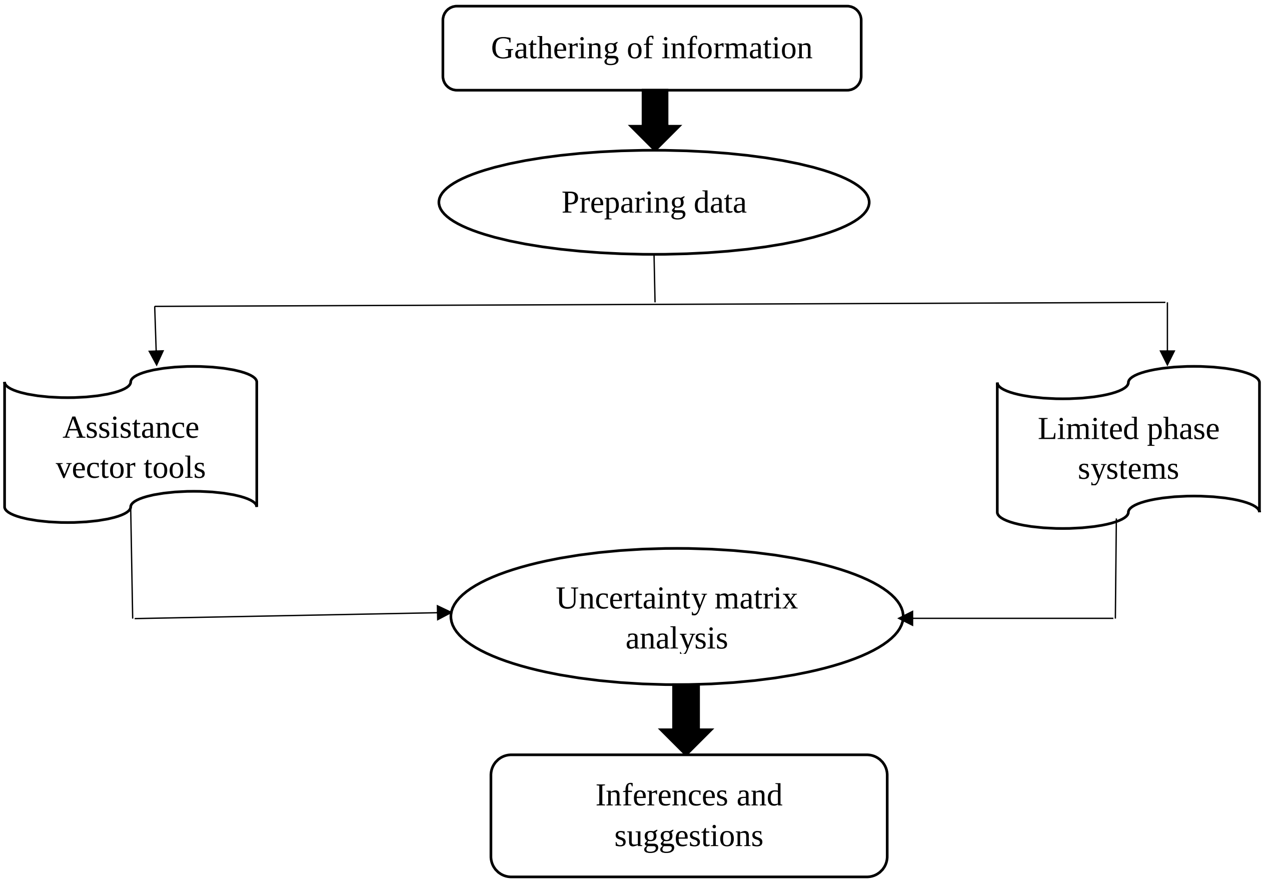

Figure 1. Methodical technique for analyzing

Together [4] the training and testing sets of user CDRs are entered into our initial ML module for detection. To get the features needed by the technique, this module uses feature extraction to extract two data sets. Several data sets—which we refer to as the identified feature set and the unlabeled feature set, respectively—are used to extract the user's feature set following labeling and feature extraction. The primary distinction between both is that the individual who signs up from the latter will be evaluated for deception, whilst the user from the former will be accepted as either a valid or fraudulent person. Next, using the marked collection of features as a basis, we employ the SVM to create a reliable binary classification algorithm.

This format will be used for the remainder of the paper of this study presented in Section 2. In Section 3, vector processing is defined, and automatic call detail record is demonstrated. We go into great depth about how learning through performance evaluation metrics is used to accomplish the great deal of the CDR using in Section 4. The topic is concluded, and future directions are outlined in Section 5.

2. RELATED WORKS

Telecom fraud has grown to be a significant issue, particularly in emerging nations like China. Currently, it may be very challenging to coordinate several agencies to fully avoid fraud. In this work, we examine the identification of big transactions sent by victims who have been duped by scammers at the bank that received them [5]. To determine the likelihood that a large transfer is forged, we suggest a novel generative adversarial network-based model. If the probability is higher than a specific limit, the institution can take the necessary precautions to keep possible scammers from stealing the money. Accomplish properly learn the intricate statistical interaction among the characteristics of the input, the deductive model uses a sophisticated blurring auto-encoding.

The perpetrators are causing a significant loss of revenues for the global telecom industry today [6]. Managers must find alternative ways to employ mathematical instruments and artificial neural network techniques to identify the reason beforehand and respond immediately to get around these kind of company hazards and maintain market share. The lack of understanding of ML and methods to be applied to such information is one factor contributing to the underutilization of this capability. Many industries, including technology, health care facilities, banks, insurance companies, and retail stores, are already utilizing artificial neural networks to gather business insights about their customers and, as a result, suggest or enhance their experiences, even though the technology is still in its infancy.

It is apparent that consumers and telecom carriers are losing a lot of money to fraud, and researchers are generally concerned about identifying and detecting scammers [7]. To combat fraudulent behavior, researchers have put up a number of remedies. Due to the reason that scammers often shift between several cellular providers, those techniques may become less successful in detecting scam. Furthermore, because of the dearth of actual data, researchers are forced to conduct simulations in a virtual environment, which weakens the persuasiveness of their models and findings. Using collaboration amongst mobile telecommunications operators, we presented a revolutionary approach in our earlier research that is very accurate and secure. Researchers will confirm it in an everyday context with actual CDR data in this publication.

It is feasible to detect fake subscribers to broadband telephone services by examining their setup and usage locations [8]. However, it takes a lot of time and cash to examine every home client in a company such as TCI, which has hundreds of thousands of residential customers, in order to identify all fake customers. Therefore, it is quite demanding to reduce the quantity of consumers to be verified. The purpose of the research is to identify trends in the behavior of home and business subscribers by analyzing their CDR and bill data. This will help distinguish between residential subscriptions, which behave similarly to fraudulent clients. Our goal is to identify the actual subscription type as accurately as possible. The depletion of a significant portion of telecommunications revenue can be avoided by discovering membership scams.

Deception has been around since the beginning of time, and yet it may take on an infinite number of forms and new tactics from con artists [9]. Fraud in telecommunications, credit card, internet, electronic cash machines, insurance, healthcare, and cash laundering are just a few of the numerous sectors where it occurs. Interception of internet connections or systems. The procedure for identifying scammers is typically identical across the aforementioned regions. Deception and loss of revenue are not a single thing. Waste of revenue due to operational or errors in technology is referred to as revenue leakage. When the causes are identified, which is typically done by implementing new inspection controls or improving construction processes, the losses may typically be recovered.

The problem of fraud has long plagued telecommunications firms [10]. In addition to the monetary losses brought on by fraudulent activity, businesses who are unable to stop these practices will also lose clients. Fortunately, fraud can be prevented by creating automated and adaptive systems. Each year, telecommunications fraud causes businesses to suffer enormous financial losses. It is difficult to figure out and declare the amount that has been lost as a result of fraud in the cellphone industry because some businesses would rather keep it a secret to preserve their brand. Additionally, telecom firms do not always identify fraud, and it is unclear how effective their verification methods are.

The Legislative stochastic framework for user profiling is a key component of our methodology. While it was first suggested to use LDA towards identifying fraud in telecoms [11], our study differs because we identify entire false accounts rather than just individual illegal calls. Thus, we use an alternative classification algorithm with the original automatic threshold-setting technique. This method can also be applied to real-world fraud detection issues. Additionally, the training phase must be completed for the strategy suggested, whereas our method is not limited by this requirement. An illustration of a stochastic mixed model is the LDA system. A model in which specific probability distributions are combined.

The quick growth of the telecom industry has additionally resulted in a rise in fraud, which damages the image of companies and revenue [12]. This research therefore suggests a novel telecom fraud detection methodology that is based on user behavior aberrations demonstrated by time-varying signatures. Given that these variations resemble well-known frauds, a list of potential suspects has been compiled and forwarded to fraud specialists for approval. Using the Map Reducing parallelism model, which offers simplicity and adaptability for large-scale applications, the suggested design was created. Ultimately, the model was used on an IT company's call detail information.

According to modern technology advances, fraud in telecoms is growing significantly, costing billions of dollars annually on a global scale. Preventive solutions are the most effective means of lowering fraud [13]. The majority of the time, fraudsters will find a method to get around the safety protocols of a business since they are adaptable and constantly looking for novel methods to commit fraud. The limitations of current strategies and techniques to detect and prevent fraud in today's telecom corporations are evaluated in this article, which also reveals some of the means that fraud is utilized against enterprises. We also provide a signature-based data mining profiling method that was created for a legitimate mobile telecoms service operator and incorporated into one of its active Fraud Detection Units.

3. METHODS AND MATERIALS

3.1 Troubleshooting vector processing machines

SVM have demonstrated outstanding adaptation accuracy across a variety of categorization issues in more recent times. To maximize the amount of difference between both positive and negative cases, SVM looks for a linear optimum at the hyperplane in binary classification situations. To do this, a quadratic optimization problem is solved, where the only important variables are support vectors, or the data points nearest to the ideal hyper-plane. But in reality, the data are frequently not linearly separable. The input space is converted via an unpredictable translation into a feature space with additional dimensions by employing a kernel function called SVM to forecast attrition subscription-services programs, hence improving the viability of the straight segregation. Figure 2 show their findings demonstrate that SVM performs well in approximation when used with distorted marketing information.

They looked into how crucial parameter selection was to SVM's performance when compared to random forests and LR used neural networks and SVM for detecting fraud with credit cards. Their findings demonstrate the efficacy of both approaches [14]. They also demonstrated that SVM may perform better than NN provided the data recordings are small. SVM ensembles for telecom subscription fraud that use either bagging or enhancing with aggregation techniques. Many electronic databases contain unprocessed, noisy, and inaccurate unprocessed information. Data in a format unsuitable for data mining models, missing values, outliers, and redundant or outdated characteristics could all be present in the databases. Therefore, preprocessing data can assist enhance the accuracy of the information and the mining outcomes as well as the extraction process's simplicity of use as well as effectiveness in terms of quality, cost, and time.

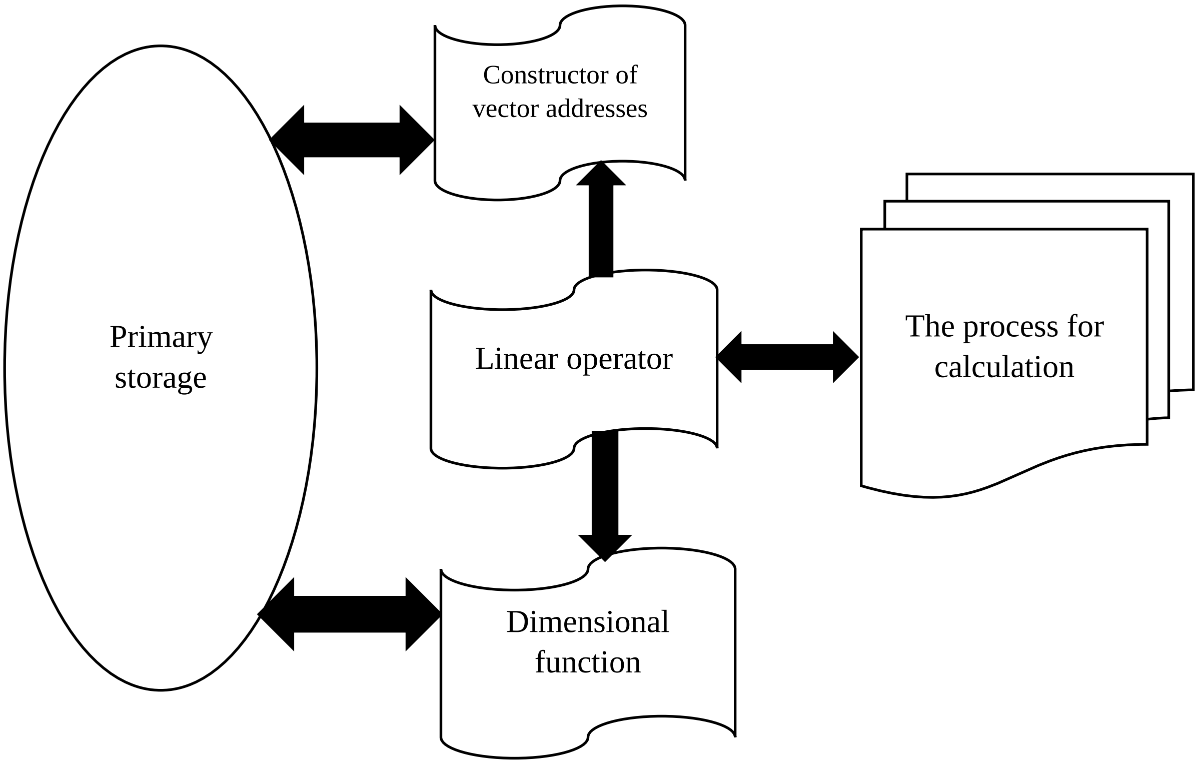

Figure 2. Linear vector analyzing device

3.2 CDR-Oriented Deception Identification

This category contains relevant works about CDR-based Scam Prevention [15]. The previous works' methodologies are categorized as both deep learning and conventional machine learning. In particular, investigation involving constructing elements is the subject of the approach provided. To categorize bogus addresses, some of the most user-friendly techniques are generating characteristics and using traditional ML techniques. There have been numerous attempts to address fraud detection issues using various features and AI methods.

3.2.1 Standard neural networks

Several studies have concentrated on the application of SVM. SVM's primary goal is to remove as much information as possible from the data. Finding an ideal hyperplane with the greatest margin is the aim of SVM. It can resolve nonlinear issues by incorporating the kernel approach utilized by SVM with a kernel of RBF to predict fraudulent users following the extraction of CDR attributes. Statistical characteristics were retrieved and used to model under a specific temporal granularity, including the number of calls, the local area mode, the call type mode, and the calling user mode. To anticipate the scam calls, they then employed SVM using linear, poly, and RBF kernels.

The most effective approach is determined by taking into account the vote outcomes of every choice tree on data samples, which is another popular family of methods. RF addresses overfitting and outperforms a single decision tree by combining or averaging the outputs of many decision trees. Created a system of analysis based on the behavior and input of consumers. The random forest model achieved approximately 83.50% accuracy on a real-world dataset after preprocessing CDR to extract features from nine distinct granularities.

The original dataset was then rebalanced using an adaptive synthetic sampling approach. Due to its superior performance on their CDR-based fraud detection dataset over k-Nearest Neighbors and SVMs with linear and RBF kernels, the study chose RF as a baseline model. Classical data mining techniques are more computationally efficient, but their feature engineering stage is difficult and time-consuming. Additionally, they are unable to reach outstanding results due to their insufficient ability to comprehend the underlying rules of data.

Despite their enormous computational capacity, the aforementioned deep learning techniques have achieved great performance; nonetheless, the majority still require specialized knowledge to construct and extract features before feeding them into models, which is time-consuming and unable to keep up with the rapidly evolving tactics of scammers. In this field, a technique to automatically obtain attractive features with more useful information is critically needed to replace time-consuming hand-designed feature engineering.

3.2.2 A synthetic neural network for computation

Neurons are connecting links, activation mechanisms for each neuron in question, and the amount of weight assigned to every connection link make up a standard NN design. Input, output, and concealed nodes are the three types of nodes that each neuron uses to process input data. The related strength associated with the node determines how to scale about the input information to a neuron. During the NN's education phase, optimal weights are determined; however, the procedures may go on during the experimentation and decision-making stages. The initial component within a vector with n characteristics, which serves as the NN's input, displays the date and time stamp. The prejudice term is expressed by the appropriate weight vector, and the NN's prediction for the university input vector is one way to convey the summing function's result.

$$v_{k=w_K^T} X_{k=\sum_{j=1}^{27} w\ Kjxkj} \qquad (1)$$

This something like a hyperplane, the boundary of decision can be expressed as follows:

$$w_{k1}\ x_{k1} + w_{k2}x_{k2}\ + w_{km}x_{km} + b \qquad (2)$$

Calculating the actual response is done with the sigmoid function provided by:

$$y_k = Sgn(w_k^T X_k) = 0 \qquad (3)$$

With the intended output by d, the adaptation vector may be expressed as

$$w(n+1) = w(n) + \gamma[d(n) - y(n)]X(n) \qquad (4)$$

Stands our experimental procedure has made use of features for typical as well as false subscribers who supply the network with input. The algorithms' performance is assessed using the following the algorithm's speed and error are visualized using an NN confusion matrix. The matrix of miscommunication and matching matrix are used, as well, to assess NN effectiveness in the context of unsupervised and supervised learning. Other names for this kind of matrix include a mistake column or a table of contingencies.

3.2.3 Measurements for Collection Quality Productivity

The instructor collection of information came from a Chinese wireless operator. Three trillion individuals created CDRs in a rate of days, making up the entire collection of information. Every time a user phones or SMS's another individual, a CDR is created. The contacting user, the called user, the start and stop times of the phone conversation or SMS, and other information are all contained in each CDR. Using the Rocket telephone number databases, which the user marks using the mobile client implementation, we categorized the entire data set, and we think it is accurate and reliable. There were two separate sets of data one for training and one for testing. In the testing set, there are three thousand legitimate users and a thousand fraudulent users, whereas the training set has 680 bogus members and two thousand valid users.

Numerous studies and companies working in this field have already made substantial use of FSM-based modeling methods. The particular framework we employed for this endeavor is currently published in several places. A complex system made up of several elements that are representations of various classes can be realized thanks to it. Such things use passing messages for interaction. Code for FSM classes is rather simple and is based on SDL diagrams. A message handler function is mapped to each SDL branch that represents a response to an event that covers a full state transition, and states of an FSM are easily identified in the SDL diagram. Complete state transitions are assumed to be necessary actions that reflect a response to the message received and culminate in a change to the same or a different state.

4. IMPLEMENTATION AND EXPERIMENTAL RESULTS

During the ML module's evaluation, a list of users who appeared suspicious was gathered. Each point is then categorized into groups according to the contacts after we extract a set of C for anybody who seems to have been at danger of fraud. Figure 3 shows the two approaches' respective accuracy, recall, and F-Measure. The figure shows that Subudhi's method achieves a detection accuracy rate of 84.19%, recall rate of 81.31%, and F-Measure of 83.05%, whereas our method achieves a high detection accuracy rate of 93.56%, recall rate of 89.22%, and F-Measure of 91.02%. Thus, we conclude that our approach outperforms Subudhi's approach. We summarize three factors that contribute to the superior performance of our approach through analysis. Subudhi's method only chooses four characteristics, however we examine and choose nine features in the machine learning module. To enhance the effectiveness of suspicious user detection, we employ a sliding window in the machine learning module to fine-tune the time granularity.

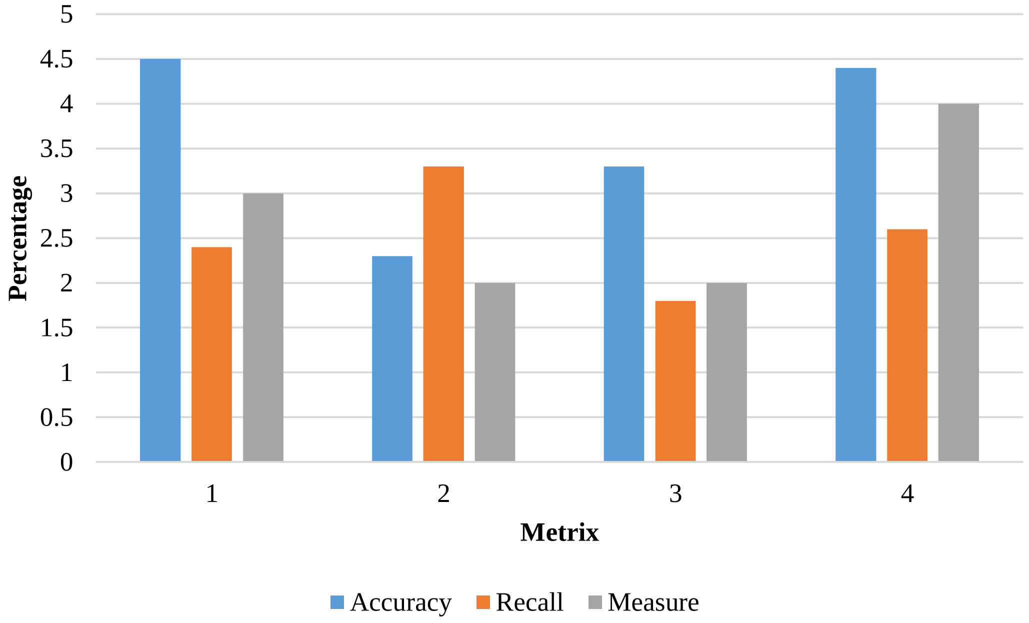

Figure 3. The accuracy score, recall and F-measure Analysis

As demonstrated in the preceding section, our approach outperforms the others in each of the three criteria. In addition, our method performs exceptionally well in terms of time efficiency. Additionally, we use Subudhi's approach to compare in the time efficiency test. The two approaches were allowed to run consecutively on four data sets, each of which included CDRs for 0.5 million, 1 million, 1.5 million, 2 million, 2.5 million, and 3 million users. Figure 4 show that illustrates the time efficiency comparison of two approaches for the four data sets. On all four data sets, the figure makes it quite evident that our approach outperforms Subudhi's. Furthermore, our method's time consumption improves consistently with the amount of the dataset, whereas Subudhi's method's time consumption increases rapidly.

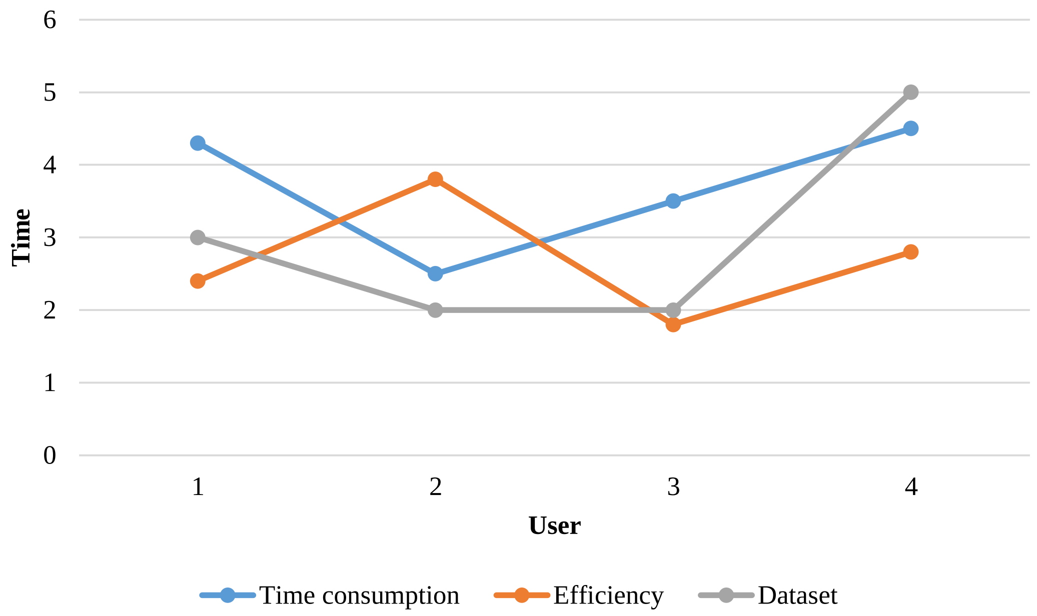

Figure 4. Evaluation of time optimization

The graphic makes it evident that our approach outperforms the other approach for each of the four data sets. Also, our method's time consumption increases consistently with the quantity of the dataset, but it climbs at an ever-increasing rate. We discovered that the ML module relates the way the user behaves characteristics on a given day to the period in question. Thus, a sliding window is used to divide a day into multiple equal time slots. Whenever the sliding window advances one step. For the sliding window approach to divide a day, windows of size N must be specified. The demographic details of users at various time points are then calculated in order to develop the feature further, better describe the person, and ultimately generate a suspect list that is more accurate.

Table 1. A list of the data used in segmentation using SVM and ML is provided.

S.no	Characterization	Value
1	Standard customers	6,965
2	Ordinary call data records	3,222,912
3	False data	30
4	Call log of fake users	20,304
5	Period for gathering data	30

Our research aims to evaluate the efficacy of Single Intelligent Box Fraud Detection by using machine learning techniques. The following steps must be taken in order to complete the procedure: recurring meeting in addition to simulator device fraud. By preparing the digital recordings, the required input properties for classification can be extracted. Information classification into three groups. Table 1 depicts applying SVM and NN to the categorizing process. Examine the outcomes by contrasting them. In that information gathering process, a quantity marker is employed to differentiate between malicious and legitimate subscribers. The results of the methods and their variants are categorized and contrasted with real data.

5. CONCLUSION

The innovative approach to scam authentication presented in the present research has shown to be successful. The ML element and the design component are two of the framework's sections. Within ML Finally, we built an algorithm by integrating the SVM method with four different types of information that we gathered. After that, we classified the test group using the algorithm to obtain a list of people who seemed suspicious. Completing carefully examining how fraudulent as well as actual customers collaborate, we build an overall scam user detection model in the template identification component then turn it into an FSM. A scam user can be identified by comparing every suspected user's CDR with FSM. The results of our examination of an actual data set demonstrate that this strategy performs better than the latest methods.

REFERENCES

[1] Li, R., Zhang, Y., Tuo, Y., & Chang, P. (2018, May). A novel method for detecting telecom fraud user. In 2018 3rd International Conference on Information Systems Engineering (ICISE) (pp. 46-50). IEEE.

[2] Kashmir, M., & Bashir, S. (2019, March). Machine learning techniques for sim box fraud detection. In 2019 International Conference on Communication Technologies (ComTech) (pp. 4-8). IEEE.

[3] Farvaresh, H., & Sepehri, M. M. (2011). A data mining framework for detecting subscription fraud in telecommunication. Engineering applications of artificial intelligence, 24(1), 182-194.

[4] Li, R., Zhang, Y., Tuo, Y., & Chang, P. (2018, May). A novel method for detecting telecom fraud users. In 2018 3rd International Conference on Information Systems Engineering (ICISE) (pp. 46-50). IEEE.

[5] Zheng, Y. J., Zhou, X. H., Sheng, W. G., Xue, Y., & Chen, S. Y. (2018). Generative adversarial network-based telecom fraud detection at the receiving bank. Neural Networks, 102, 78-86.

[6] Daka, J. C., & Nyirenda, M. (2022). Smart Mobile Telecommunication Network Fraud Detection System Using Call Traffic Pattern Analysis and Artificial Neural Network. American Journal of Intelligent Systems, 12(2), 43-50.

[7] Ruan, N., Wei, Z., & Liu, J. (2019). Cooperative Fraud detection model with privacy-preserving in real CDR Datasets. IEEE Access, 7, 115261-115272.

[8] Farvaresh, H., & Sepehri, M. M. (2011). A data mining framework for detecting subscription fraud in telecommunication. Engineering applications of artificial intelligence, 24(1), 182-194.

[9] Bramantoro, A., & Alraouji, Y. International Call Fraud Detection Systems and Techniques.

[10] Babaei, K., Chen, Z., & Maul, T. (2019). A Study of Fraud Types, Challenges and Detection Approaches in Telecommunication. Journal of Information Systems and Telecommunication, 7(4), 248-261.

[11] Olszewski, D. (2012). A probabilistic approach to fraud detection in telecommunications. Knowledge-Based Systems, 26, 246-258.

[12] Terzi, D. S., Sağıroğlu, Ş., & Kılınç, H. (2021). Telecom fraud detection with big data analytics. International Journal of Data Science, 6(3), 191-204.

[13] Lopes, J., Belo, O., & Vieira, C. (2011). Applying user signatures on fraud detection in telecommunications networks. In Advances in Data Mining. Applications and Theoretical Aspects: 11th Industrial Conference, ICDM 2011, New York, NY, USA, August 30–September 3, 2011. Proceedings 11 (pp. 286-299). Springer Berlin Heidelberg.

[14] Farvaresh, H., & Sepehri, M. M. (2011). A data mining framework for detecting subscription fraud in telecommunication. Engineering applications of artificial intelligence, 24(1), 182-194.

[15] Zhen, Z., & Gao, J. (2023). CDR2IMG: A Bridge from Text to Image in Telecommunication Fraud Detection. Comput. Syst. Sci. Eng., 47(1), 955-973.