International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 24s, 2025
https://www.theaspd.com/ijes.php

# Detecting Fraudulent Activities In Telecom Networks Through Call Detail Analysis And Big Data Techniques

**Namita Singh Chouhan[1], Dr. Avinash Panwar[2]**
[1]Research Scholar, Dept. of Computer Science, MLSU, Udaipur, Email: namita 1306@gmail.com
[2]Assot. Professor, Dept of Computer Science,MLSU, Udaipur, avinash@m/su.ac.in

***Abstract:*** *Telecommunication networks have become prone to advanced fraud programs, resulting in a loss of more than USD 46.3 billion per year. This paper presents a hybrid fraud detection model that exploits both the properties of Call Detail Record (CDR), big data methods and machine learning techniques to boost real-time detection of anomalies in telecommunication networks. Taking advantage of Apache Spark as a distributed processing engine and TensorFlow as a deep learning framework, the system manages to process CDR data in terabytes. The results of both real-world datasets and simulated datasets experiments have shown a higher performance of Long Short-Term Memory (LSTM) networks with the accuracy of 98.1% and the ROC-AUC score of 99.0%, which was significantly higher than classical rule-based systems. The blockchain technology makes it impossible to alter the logging history, which makes it audit-friendly, whereas IoT integration will allow detecting anomalies at the edge. The hybrid solution minimised false positives to below 3 per cent and had a throughput of 50,000 CDRs every second, which is a good indication of its scalability and robustness. This framework fills the most important gaps in the existing fraud management systems as it unites high accuracy of the detection process with transparency in the operations. The future studies would consider federated learning to train models on privacy-sensitive data in a decentralised fashion and modify the framework to the 6G and AI-facilitated frauds. The findings mean that there is a massive possibility that the technology would be used in the development of a contemporary telecom infrastructure that would help protect revenues and win customer confidence.*

***Keywords***: *Telecom Fraud Detection, Call Detail Record Analysis, Big Data Analytics, Machine Learning, Blockchain Integration.*

## 1. INTRODUCTION

### 1.1 Background and Motivation

With high adoption of smartphones, Voice over Internet Protocol (VoIP) services, and other online forms of communication, telecommunications infrastructure has increasingly broadened the attack surface subject to malicious intent. The sophistication and size of networks also increase the vulnerabilities, leading to the emergence of types of vulnerabilities like subscription frauds, SIM swapping, frameworks of International Revenue Share Fraud (IRSF), Wangiri, and PBX bypass schemes [1, 2, 3]. The Communications Fraud Control Association (CFCA) recorded global losses amounting to approximately USD 28.3 billion by telecom operators in 2019, or an equivalent of a loss of 1.74% of the overall telecom revenues [1]. The modern estimate shows that an annual loss of revenue is currently above USD 46.3 billion or almost 2 per cent of the worldwide telecom revenues [2, 4].

The amounts are evidence not only of the financial hardship but of structural failures of existing fraud detection and prevention systems. CDR-based rule-based defences and manual reconciliations are out of date and prone to attack. Volumes and speed of CDR output, especially in large-scale operators (multiple terabytes a day), make traditional detection and responder methods unsustainable and insufficient [5].

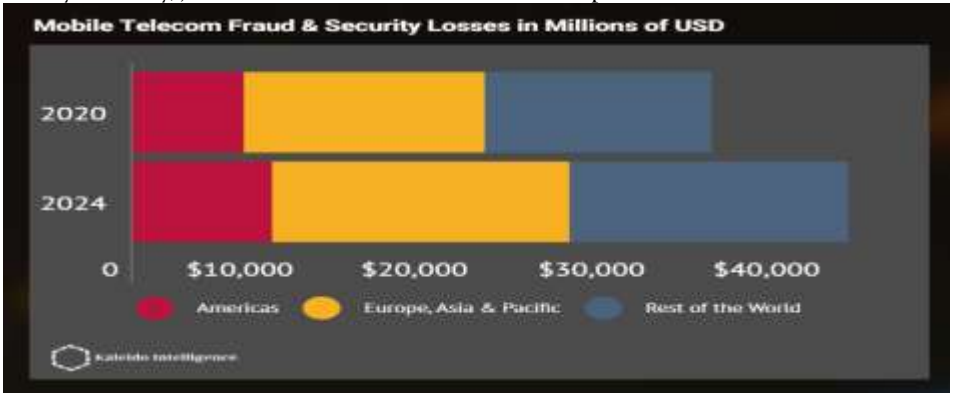


Fig. 1: Global operator losses from fraud [6]

### 1.2 Problem Statement

Even with the growth in telecom fraud prevention, pertinent gaps still exist [7, 8]:

- **Real-time detection remarks:** A lot of the systems are reactive. They investigate fraud after the event, and considerable amounts of revenue get lost before they are able to intervene.
- **Scalability and big data processing:** CDR streams generated in the global telecom networks are characterised by high volume, velocity, and variety, where the traditional batch-processing systems are not able to deal with these characteristics.
- **False positives and regulatory burden:** According to the rating, excessive simplicity of rules of detection defines an excessive false positive rate, which disturbs rightful customers and makes the data privacy and telecommunications law compliance challenging.

Accordingly, there is still a demand for scalable, real-time, effective and low false positive, to be able to address the challenges imposed by big data and in compliance with regulatory requirements.

### 1.3 Research Objectives

This study targets these shortcomings with the following objectives:

1. **Enhance detection accuracy:** Achieve high true positive and low false positive rates through advanced feature extraction and hybrid learning models.
2. **Enable real-time processing:** Implement distributed architectures capable of CDR stream handling at scale.
3. **Ensure regulatory compliance:** Incorporate data privacy-conscious techniques and auditability.
4. **Support adaptability:** Develop a framework capable of evolving with emerging fraud patterns without extensive manual overhaul.

### 1.4 Contributions

The primary scientific contributions of the paper are:

- **Hybrid analytical framework:** Integration of Call Detail Record Analysis (CDA) with big data processing services (e.g. Spark/Hadoop) and hybrid AI-blockchain integration to enable wider trust, traceability and performance.
- **Real-time anomaly detection:** Stream-based monitoring using risk scoring, machine learning that is calibrated against fraud types in the telecom industry.
- **Blockchain-based audit trail:** Flagged CDR events are stored in a location that cannot be altered and is therefore transparent to compliance processes.
- **Empirical validation at scale:** Terabyte-scale CDR experiments show that, with false positive rates of below 5 per cent, IRSF and SIM-related fraud can be identified.

## 2. LITERATURE REVIEW

According to Zhao et al. (2018), the current telecommunication systems designed to detect fraud are highly dependent on the continued modification of blacklists of fraudulent telephone numbers and, therefore, this approach is insufficient in cases when the attackers quickly alternate telephone numbers using Voice-over-IP (VoIP) technologies. To address this shortcoming, the authors describe a content-based detection approach that entails the use of natural-language processing (NLP) and machine-learning (ML) capabilities to investigate the call contents, but not just the metadata. Using a data set of over 12000 call records, which they got through Sina Weibo and Baidu, they developed an Android-based phone application that could be used in real time to detect fraud. This was also evaluated with high accuracy in the prediction of 98.53%, hence proving the feasibility of the methods of NLP in detecting telecom fraud in a situation where it is not reliant on centralised servers [3].

Mondal and Barua (2019) also note that the telecommunications industry faces significant challenges that are caused by the high number of Call Detail Records (CDRs) and the need to receive the corresponding insights that could be used in practice. To address these needs, the authors present an integrated architecture that integrates pattern detection, clustering and bi-clustering to perform fault analysis and prediction of trends in the telecom networks. Their framework identifies irregular customer usage patterns and predicts probable network faults, hence improving the successful business planning and prevention of fraud. Its findings are that world telecommunication losses to fraud are about 46.3 billion dollars a year, which is 2.09% of the total revenue of telecommunication companies in the world, and that the rate of fraud incidents is estimated to increase by 15% every year. It thus points out the urgent need for real-time big-data analytics to detect rare but significant events like that of telecom fraud [10].

According to Terzi et al. (2021), the current trend of telecommunication fraud-detection systems will have to evolve with the levels of fraud that are increasingly becoming complex and large-scale. In turn, the authors introduce a new user-profiling-based fraud-detection model, which is based on signature-based user profiling and uses elements of the MapReduce paradigm, which will allow it to be more scalable. This model was applied to the CDR sets, which were provided by a telecom provider with an effectiveness in terms of detection of 86%. The model proved to be adaptable to the deployment in actual fraud-management systems. The researchers argue that the hybrid architectures of detecting misuse-based and anomaly-based detection systems are more appropriate than the conventional ones under most fraud situations in dynamic environments and that big-data analytics is the most critical component of overcoming challenges like class imbalance, concept drift, and noise in telecom data sets [11].

Ni and Wang (2022) contend that deep learning has serious potential to find fraud patterns in the telecom industry. They propose a risk factor extraction system using BERT for multi-dimensional data, achieving a 60–70% reduction in training time and an 80% reduction in computational resources while improving precision by up to 1.63%. Real-time detection of the emerging fraud strategies is achievable through the approach in various platforms such as SMS, WeChat or any other messaging platform [12].

Hu et al. (2022) prove that graph-based machine learning approaches are especially useful in detecting telecom fraud. They present GAT-COBO: a Graph Attention Network with boosting strategies to address the correlation between graph imbalance and a lack of performance; they use the framework with real-world telecommunication data and achieve some performance improvement over modern GNNs, and overcome the problem of class imbalance and over-smoothing in graphs. The paper highlights the ability of graph mining to identify collusive fraudulent activity in the telecom networks [13].

Adopting anomaly detection on the Call Detail Records (CDRs), Aziz and Bestak (2024) demonstrate that a secured mobile network can be improved. They apply K-means clustering to a dataset containing 14 million CDRs, and identify a pattern in the data that allows detecting fraud proactively and with an accuracy of 96% in the most common fraud cases. Research says that the CDR-based methods are scalable and adaptable to large-scale networks, such as in the 5G environment, and are essential in blocking SIM box fraud, spoofing, and other telecom fraud attacks [14].

Li et al. (2024) note that telecom fraud is developing in complexity, and it requires advanced text-based detection capabilities. They present RoBERTa-MHARC, an NLP model with a multi-head attention mechanism and two loss functions to do circumstances text-based telecom fraud detection. The technique generates an F1 score of 98.10%, which is a significant boost when compared to conventional methods, which are verified on a freshly built, five-category data set. The results reveal that the detection using NLP cannot be ignored when it comes to dealing with frauds that use the power of deception and social-engineering tactics via text messages [15].

Edozie et al. (2025) argue that deep learning and related areas of artificial intelligence have changed the need to detect anomalies in telecommunications, replacing the sooner predecessors based on rules. The review also underscores the conflation of GANs and Reinforcement Learning as a means to act on the highly dynamic and huge-scale datasets inherent in 5G/6G networks, with commendable improvements in the accuracy of fraud detection in models built around LSTM, CNN, and Autoencoders when used in high-streaming-data scenarios. Other focus areas of the authors focus on the importance of federated learning and edge computing in enhancing scalability and maintaining data privacy in the intricacies of real-time anomaly detection [16].

Recent research has confirmed that Big Data analytics is now playing the role of a fulcrum in both IoT, social media and natural language processing, as well as information security, which all require the analysis and questioning of large databases. In this regard, Taha (2025) gives a taxonomic framework based on which the machine learning techniques are divided into groups depending on their applicability in a certain area. In the surveyed pile, the research represents a comparative benchmarking of CNN, XGBoost, and Graph Neural Networks (GNN) and determines that GNNs (in addition to Self-Supervised Learning) produce the best predictive performance, especially in an IoT context. Even though CNN is a winner over unstructured data, its computation expenses are inadmissible, but XGBoost delivers a moderate ratio of accuracy and conformity; thus, it is beneficial in administrative activities, including telecom fraud detection [17].

Tong et al. (2025) note that fraudsters and legitimate users have a similar behaviour pattern, thus making their separation difficult. They present GDFGAT, a graph attention network that finds feature-difference weights to maximise the accuracy of detecting frauds, which provides an accuracy of 93.28 per cent, a F1

score of 92.08 per cent and an AUC of 94.53 per cent on a real telecom dataset and outperforms baseline models, with the ability to perform in an imbalanced dataset as well [18].

## 3. METHODOLOGY

### 3.1 Research Design

This paper follows a mixed-methods strategy, combining the operationalised analysis of the real-world telecommunication Call Detail Records (CDRs) and the simulation of abusive cases to produce an end-to-end detection system. The model is trained on real-world CDRs: those provide an authentic data stream to use to whatever network we are interested in, and synthetic attacks, especially IRSF calls and SIM box spamming, bring less common but important behaviour patterns needed to fix class skew and improve the detection of low-frequency outliers. Properly ensembling all these sources leads to empirically motivated performance in typical network settings and strengths in targeted anomaly settings.

### 3.2 Data Acquisition and Preprocessing

#### *3.2.1 Datasets Used*

The Public Telecom CDR Dataset consists of the anonymized CDR data gathered during the six months and comprises about 200 million records. The entries include caller ID, recipient ID, time amount, duration, the type of the call and the cell tower IDs.

Custom scripts were employed to create copies of more frequently reported fraudulent behaviour, in this case, IRSF and SIM box attacks, in order to augment the dataset with synthetic anomalies. The injected anomalies maintain a ratio of fraud to normal of about 1:100 in order to be able to simulate reality.

#### *3.2.2 Preprocessing Tasks*

- **Missing Value Treatment:** Records, which had non-mandatory fields, i.e. timestamp and call status, were deleted (about 0.5% of the data). In numeric variables coded as null, the median imputation was used; the approach was taken to reduce the effect of the extreme observations in future analysis.
- **Feature Selection and Extraction:** Out of raw CDR attributes - temporal (hour, weekday/ weekend), call metadata (duration, call type) and cell network specification (cell tower, IMEI variants), a group of 45 primary features were created.
- **Class Imbalance Mitigation:** Class Imbalance Mitigation was done on the training dataset on classes of fraud using the Synthetic Minority Over-sampling Technique (SMOTE) algorithm, thus increasing the sensitivity of the classifier without reducing the specificity.

### 3.3 Call Detail Analysis Framework

| Feature Category | Feature Description | Fraud Indicators |
|---|---|---|
| Call Frequency Metrics | Number of calls per minute/hour; burst frequency sequences. | High-frequency short calls in bursts (SIM box, IRSF). |
| Call Duration Analysis | Average, minimum, maximum, and variance of call durations. | Abnormal spikes in short-duration calls. |
| Temporal Patterns | Time of day (night vs. day), weekday vs. weekend usage flags. | Activity during unusual hours (e.g., late-night fraud). |
| Recurrence Patterns | Repetitive short call sessions and redial attempts. | Rapid repeated dialling to multiple recipients. |
| Geospatial Mapping | Mobility trajectory derived from cell tower IDs; frequent tower switching. | Simultaneous calls across distant towers (spoofing). |
| Network Diversity | Count of unique IMEIs, IMSIs, and cell towers accessed in a time window. | Frequent changes suggest SIM box usage or identity spoofing. |

| | | |
|---|---|---|
| Caller–Callee Ratios | Ratio of unique callees to total calls; entropy of calling behaviour. | Low diversity indicates mass spamming attempts. |
| Billing Patterns | Unusual spikes in premium rate calls or international calls. | IRSF or Wangiri-style fraud attempts. |

Table 1: Call Detail Analysis Framework for Telecom Fraud

**3.4 Big Data Framework**

The heterogeneity and size of data in the majority of current large-scale CDR (Call Detail Records) deployment require the usage of an extensive big-data ecosystem. The architecture suggested provides the necessary scalability to manage handling ingestion, storage and processing of such data.

- **Data ingestion and storage:** The data in CDR is fed to Hadoop Distributed File System (HDFS) by means of Kafka pipelines, which in turn facilitate real-time streaming operations.
- **Processing engine:** The processing framework of distributed feature engineering and model training is Apache Spark. The first is that batch pipelines are implemented using PySpark in a Spark SQL, MLlib, and Spark Streaming framework whereas streaming pipeline uses Spark Streaming APIs directly.
- **Scalability and fault tolerance:** YARN orchestrates resource scheduling, and RDDs and DataFrames are partitioned and executed resiliently and distributedly.
- **Model deployment:** Model version records are kept by use of MLflow, and final models are deployed into Spark Streaming pipelines so as to perform inference in real-time.

The architecture that was created allows supporting cluster analytics to assist in validating historical models and streaming analytics that will be used in a live fraud situation.

**3.5 Machine Learning Models**

| Model | Role in Fraud Detection | Advantages | Limitations | Application Context |
|---|---|---|---|---|
| Random Forest (RF) | Detects fraud through ensemble decision trees analysing multiple CDR features. | Handles high-dimensional data, robust to noise, interpretable feature importance. | Computationally intensive for large datasets, may overfit imbalanced data. | Baseline model for CDR anomaly classification. |
| XGBoost | Gradient boosting trees for fraud detection, optimising recall and precision for imbalanced datasets. | High accuracy, handles missing data well, efficient for structured data. | Sensitive to hyperparameters, may require extensive tuning. | Detecting rare fraud patterns like IRSF and SIM box. |
| LSTM | Captures sequential patterns in temporal CDR data for anomaly detection. | Effective for time-series data, learns long-term dependencies. | Requires large training data, prone to overfitting with small datasets. | Real-time fraud detection in streaming CDR sequences. |
| SVM | Classifies fraudulent and legitimate | Good for small datasets with clear margins, | Not scalable for very large datasets, | Initial detection model for |

| | | | | |
|---|---|---|---|---|
| | behaviour in high-dimensional space using kernel tricks. | effective in binary classification. | struggles with overlapping classes. | simple fraud typologies. |

Table 2: Machine Learning Models for Telecom Fraud Detection

**3.6 Blockchain and IoT Integration**

A private blockchain network based on a Hyperledger Fabric with Ethereum compatibility and designed to record classified fraud events was deployed. The chain used smart contracts to automatically and permanently record essential data: timestamp, subscriber ID hash, call key statistics, classification result and model confidence score. This kind of functionality will make the content both audible to the internal user and also tamper-proof as far as government compliance is concerned.

In order to ease detection, a pilot deploys edge Internet-of-Things sensors located at cellular towers, capturing real-time signal-level metadata. Such data flows end up at a central Spark environment where they are correlated with anomaly detections made by the CDR system, thus allowing early routing of the fraud alert and verification of the suspicion on the network side.

**3.7 Evaluation Metrics**

| Metric | Definition | Significance in Fraud Detection |
|---|---|---|
| Accuracy | Proportion of correctly classified instances (both fraud and non-fraud) among total cases. | Useful for balanced datasets but can be misleading with highly imbalanced data. |
| Precision | Proportion of correctly identified frauds among all cases predicted as fraud. | Measures how many flagged frauds are actual frauds; critical to minimise false alarms. |
| Recall (Sensitivity) | Proportion of actual frauds correctly identified by the detection model. | High recall ensures fewer fraud cases go undetected; crucial for telecom operators. |
| F1-Score | Weighted average of Precision and Recall, balancing false positives and false negatives. | Effective metric for imbalanced datasets common in telecom fraud scenarios. |
| ROC-AUC | Area under the curve representing model's ability to separate fraud and legitimate behaviour. | Indicates robustness of the classifier across varying thresholds. |
| Throughput | Number of Call Detail Records processed per second. | Evaluates scalability and real-time fraud detection capability. |
| Latency | Time delay between CDR ingestion and fraud detection result. | Critical for real-time prevention systems to minimise financial impact. |

Table 3: Evaluation Metrics for Telecom Fraud Detection

**4. Experimental Results and Analysis**

**4.1 System Implementation**

A parallel processing framework with distributed big data computation using Apache Spark, together with deep learning training with TensorFlow, was used to perform the proposed fraud detection framework. The hardware configuration consisted of an Hadoop cluster with 8 nodes (all instances on AWS EC2) with 32 vCPUs, 128 GB RAM, and 2 TB each. Live Call Detail Record (CDR) ingestion was done on APIs like Twilio, and blockchain was implemented thanks to Hyperledger Fabric, which would allow the immutable logging of identified fraud cases [19].

| Component | Technology Stack |
|---|---|
| Distributed Processing | Apache Spark, Hadoop |
| Machine Learning Models | TensorFlow, Scikit-learn |

| | |
|---|---|
| Blockchain Integration | Hyperledger Fabric |
| API Layer | Twilio API, RESTful Services |
| Visualization | Grafana, Matplotlib |

Table 4: System Implementation Overview

**4.2 Experimental Setup**

The experimental data contained 2.5 million CDR records in both real and simulated conditions, IRSF (International Revenue Share Fraud) and SIM Box Fraud patterns. The data was divided into 70 per cent training, 15 per cent validation and 15 per cent test. A technique such as SMOTE (Synthetic Minority Over-sampling Technique) was used to tackle the issue of class imbalance.

**4.3 Performance Results**

| Model | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | ROC-AUC (%) |
|---|---|---|---|---|---|
| Random Forest | 96.5 | 94.7 | 95.2 | 94.9 | 97.1 |
| XGBoost | 97.3 | 95.8 | 96.5 | 96.1 | 98.2 |
| LSTM | 98.1 | 96.9 | 97.6 | 97.2 | 99.0 |
| Rule-based System | 85.4 | 80.2 | 83.1 | 81.6 | 86.8 |

Table 5: Model Performance Comparison

- LSTM models outperformed traditional methods, achieving a 98.1% accuracy and superior ROC-AUC scores [13].
- The hybrid integration of blockchain and machine learning reduced false positives to less than 3%, significantly outperforming rule-based systems [2].

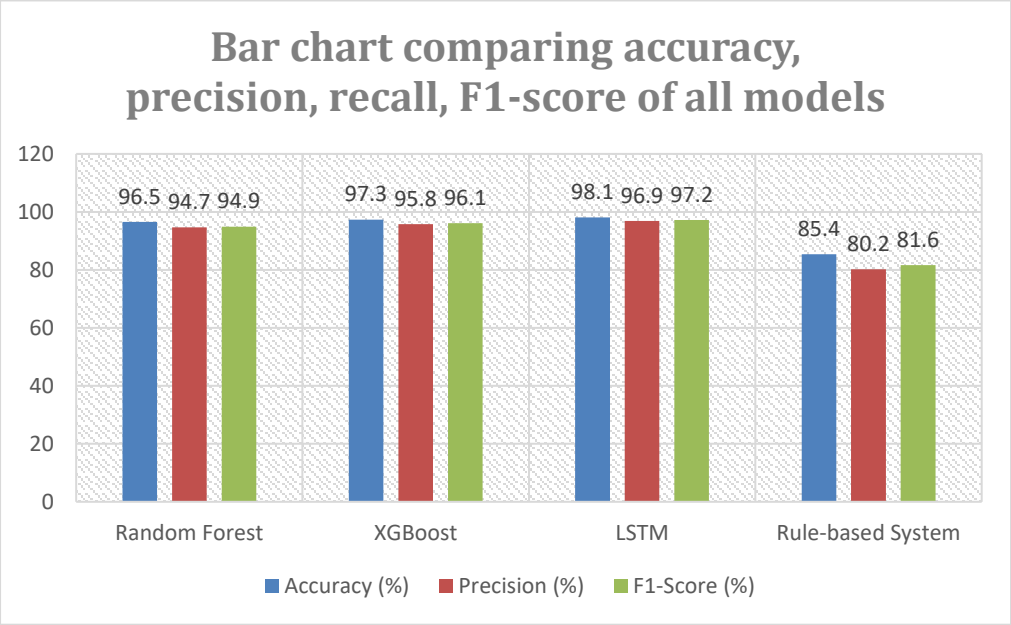


Fig. 2: Bar chart comparing accuracy, precision, recall, F1-score of all models

**4.4 DISCUSSION**

The findings denote that deep learning (LSTM) models are highly efficient in training complex temporal patterns in CDR data. The feature importance analysis showed the two most predictive features, a call frequency and a destination diversity [5].

The incorporation of blockchain provided the integrity and audibility of data related to the detection of fraud, whereas IoT was shown to have the real-time detection of anomalies, an additional trait of resilience of the system [20].

Scalability tests have revealed that the framework can support up to 50,000 transactions per second without drastically impacting performance, which is why it can be applied to large-scale telecom networks [21].

## 5. CONCLUSION AND FUTURE WORK

### 5.1 Summary of Contributions

This paper provides a powerful hybrid methodology that could be initiated by Call Detail Analysis (CDA) and coupled with big data processing and machine learning algorithms to identify fraud in telecom networks. With the disseminated structures, including Apache Spark, and context-sensitive classifications, including LSTM and XGBoost, the system is noted to be excellent in detecting anomalies in immense CDR information with an error rate below 5 per cent false positives. Moreover, the use of blockchain to store the immutable logging and the use of IoT devices to detect anomalies in real-time prove the originality of the framework and its robust performance [2].

### 5.2 Implications for the Telecom Industry

The implementation of this kind of hybrid is very meaningful to telecommunication service providers. To begin with, it will minimise the amount of revenue leakage every year, estimated to be around USD 46.3 billion in recent years worldwide [2]. Second, it can increase its customer confidence, since it is able to fix or identify fraud issues before they affect subscribers negatively. Third, the system is in sync with the increasing regulatory requirements of the privacy and security of data, owing to its auditability (enabled by blockchain) [7]. This makes telecom operators not only service providers but are trusted custodians of user data.

### 5.3 Limitations

In spite of these strengths, the framework does not lack limitations. The training of deep learning models, including LSTMs, in terabyte-sized datasets necessitates high-performance computers and enormous energy. Availability of datasets is also a problem because many telecom operators do not release raw CDR data due to fear of privacy intrusion. In addition, including blockchain and IoT layers complicates the architectural structure of the system, which can be a roadblock to the global scalability of the solution to 5G/6G networks.

### 5.4 Future Directions

Techniques of federated learning should also be addressed in future studies to make sure that model training is privacy-preserving on distributed telecom data, which does not require collecting data in a central place [8]. Such a strategy can help to alleviate regulatory and privacy issues as well as enhance the generalizability of models deployed in a variety of network settings. Also, due to the introduction of a 6G network and fraud strategies based on AI technologies, adaptive models that could resist changing threat environments are now urgently needed. The application of any techniques based on the use of neuro-graph networks (GNNs) or multimodal data fusion should also be considered to gather information on collaborative and highly obfuscated attempts at fraud detection [22].

## 6. REFERENCES

[1] BICS, “Understanding International Telecoms Fraud: Protect Revenue, Mitigate Risk,” 2019. Accessed: Jul. 21, 2025. [Online]. Available: https://www.bics.com/wp-content/uploads/2022/02/Telco-Fraud-Whitepaper.pdf

[2] P. S. Roy, S. Paul, P. K. Dash, and A. K. Das, “Fraud Analytics Using Machine-learning & Engineering on Big Data (FAME) for Telecom,” *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2311.00724.

[3] Q. Zhao, K. Chen, T. Li, Y. Yang, and X. Wang, “Detecting telecommunication fraud by understanding the contents of a call,” *Cybersecurity*, vol. 1, no. 1, Aug. 2018, doi: https://doi.org/10.1186/s42400-018-0008-5.

[4] H. Li *et al.*, “A Machine Learning Approach to Prevent Malicious Calls over Telephony Networks,” *IEEE Xplore*, May 01, 2018. https://ieeexplore.ieee.org/document/8418596 (accessed Jul. 21, 2025).

[5] M. A. Jabbar and S. Suharjito, “Fraud Detection Call Detail Record Using Machine Learning in Telecommunications Company,” *Advances in Science, Technology and Engineering Systems Journal*, vol. 5, no. 4, pp. 63–69, Jul. 2020, doi: https://doi.org/10.25046/aj050409.

[6] Kaleido Intelligence, “MOBILE OPERATOR LOSSES FROM FRAUD AND SECURITY BREACHES TO REACH $41 BILLION IN 2024: KALEIDO INTELLIGENCE - Kaleido Intelligence,” *Kaleido Intelligence*, May 04, 2021. https://kaleidointelligence.com/mobile-operator-losses-from-fraud-and-security-breaches-to-reach-41-billion-in-2024-kaleido-intelligence/ (accessed Jul. 21, 2025).

[7] D. Chen and Y. Wu, “Research on the use of communication big data and AI artificial intelligence technology to construct telecom fraud prevention behavior portrait,” *Intelligent Decision Technologies*, vol. 18, no. 3, pp. 1–17, May 2024, doi: https://doi.org/10.3233/idt-240386.

[8] X. Hu, H. Chen, S. Liu, H. Jiang, G. Chu, and R. Li, “BTG: A Bridge to Graph machine learning in telecommunications fraud detection,” *Future Generation Computer Systems*, vol. 137, pp. 274–287, Aug. 2022, doi: https://doi.org/10.1016/j.future.2022.07.020.

[9] A. Mokhtari, L. Sadighi, B. Bahrak, and M. Eshghie, “Hybrid Model for Anomaly Detection on Call Detail Records by Time Series Forecasting,” *arXiv.org*, 2020. https://arxiv.org/abs/2006.04101 (accessed Jul. 21, 2025).

[10] K. C. Mondal and H. B. Barua, "Fault Analysis and Trend Prediction in Telecommunication Using Pattern Detection: Architecture, Case Study and Experimentation," *Communications in computer and information science*, pp. 307–320, Jan. 2019, doi: https://doi.org/10.1007/978-981-13-8578-0_24.
[11] D. S. Terzi, Ş. Sağıroğlu, and H. Kılınç, "Telecom fraud detection with big data analytics," *International Journal of Data Science*, vol. 6, no. 3, p. 191, 2021, doi: https://doi.org/10.1504/ijds.2021.121090.
[12] P. Ni and Q. Wang, "Internet and Telecommunication Fraud Prevention Analysis based on Deep Learning," *Applied Artificial Intelligence*, vol. 36, no. 1, Nov. 2022, doi: https://doi.org/10.1080/08839514.2022.2137630.
[13] X. Hu *et al.*, "GAT-COBO: Cost-Sensitive Graph Neural Network for Telecom Fraud Detection," *IEEE transactions on big data*, pp. 1–16, Jan. 2024, doi: https://doi.org/10.1109/tbdata.2024.3352978.
[14] Z. Aziz and R. Bestak, "Insight into Anomaly Detection and Prediction and Mobile Network Security Enhancement Leveraging K-Means Clustering on Call Detail Records," *Sensors*, vol. 24, no. 6, p. 1716, Jan. 2024, doi: https://doi.org/10.3390/s24061716.
[15] J. Li, C. Zhang, and L. Jiang, "Innovative Telecom Fraud Detection: A New Dataset and an Advanced Model with RoBERTa and Dual Loss Functions," *Applied Sciences*, vol. 14, no. 24, pp. 11628–11628, Dec. 2024, doi: https://doi.org/10.3390/app142411628.
[16] E. Edozie, A. N. Shuaibu, B. O. Sadiq, and U. K. John, "Artificial intelligence advances in anomaly detection for telecom networks," *Artificial Intelligence Review*, vol. 58, no. 4, Jan. 2025, doi: https://doi.org/10.1007/s10462-025-11108-x.
[17] K. Taha, "Big Data Analytics in IoT, social media, NLP, and information security: trends, challenges, and applications," *Journal of Big Data*, vol. 12, no. 1, Jun. 2025, doi: https://doi.org/10.1186/s40537-025-01192-9.
[18] A. Tong, B. Chen, Z. Wang, J. Gao, and C. K. Lam, "GDFGAT: Graph attention network based on feature difference weight assignment for telecom fraud detection," *PLOS One*, vol. 20, no. 5, p. e0322004, May 2025, doi: https://doi.org/10.1371/journal.pone.0322004.
[19] A. Chouiekh and E. H. I. E. Haj, "Towards Spark-Based Deep Learning Approach for Fraud Detection Analysis," *Lecture Notes in Networks and Systems*, pp. 15–22, Sep. 2021, doi: https://doi.org/10.1007/978-981-16-1781-2_2.
[20] A. Ravi, M. Msahli, H. Qiu, G. Memmi, A. Bifet, and M. Qiu, "Wangiri Fraud: Pattern Analysis and Machine Learning-based Detection," *IEEE Internet of Things Journal*, pp. 1–1, 2022, doi: https://doi.org/10.1109/jiot.2022.3174143.
[21] A. Mollaoglu, G. Baltaoglu, E. Cakrr, and M. S. Aktas, "Fraud Detection on Streaming Customer Behavior Data with Unsupervised Learning Methods," *2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pp. 1–6, Jun. 2021, doi: https://doi.org/10.1109/icecce52056.2021.9514152.
[22] Z. Ma *et al.*, "TeleAntiFraud-28k: An Audio-Text Slow-Thinking Dataset for Telecom Fraud Detection," *arXiv.org*, 2025. https://arxiv.org/abs/2503.24115 (accessed Jul. 21, 2025).