A Scalable Hybrid Approach to Detecting Fraud with Machine Learning

1st Syrine Ben Abid
Chapter Gen AI & LLM Services
T-Systems International GmbH
Darmstadt, Germany
syrine.ben-abid@t-systems.com

2nd Bernhard Fuchs
Chapter Engineering & Data Analytics
Deutsche Telekom Security GmbH
Bonn, Germamy

3rd Günter Haberkorn
Chapter Fraud & Abuse Management
Deutsche Telekom Security GmbH
Nürnberg, Germany

4th Kathrin Jepsen
Chapter Product Management
Deutsche Telekom AG
Berlin, Germany

5th Alexander Krampetz
Chapter Fraud & Abuse Management
Deutsche Telekom Security GmbH
Stuttgart, Germany

6th Detlev Matthes
Chapter SWE Data
Deutsche Telekom IT GmbH
Berlin, Germany

***Abstract*—Rule-based fraud detection struggles with evolving tactics and fixed thresholds, limiting its effectiveness. This study focuses on detecting SMS fraud using real Call Detail Records and introduces a hybrid machine learning framework combining supervised and unsupervised techniques. XGBoost, trained on expert-validated rule-based labels, benchmarks detection, while an autoencoder, Isolation Forest and DBSCAN identify anomalies and structure fraud patterns. To refine detection, the Fraud Relevance Index integrates model outputs through a weighted sum, prioritizing high-risk cases. The proposed framework improves fraud detection accuracy, reduces reliance on manual labeling and adapts to emerging fraud scenarios.**

***Index Terms*—Telecom Fraud Detection, Machine Learning, Supervised Learning, Unsupervised Learning, Anomaly Detection**

I. Introduction

Fraud detection in telecommunications is a continuous challenge, as fraudsters adapt to evade detection. Traditional rule-based systems, relying on static thresholds, struggle to detect emerging fraud patterns, leading to both false positives (misclassified legitimate users) and false negatives (missed fraud cases).

This work focuses on Short Message Service (SMS) fraud, which manifests in various forms including phishing scams, SMS spamming, malware distribution, and bulk fraud messaging targeting users with fake promotions or illicit services.

Machine learning (ML) provides a scalable alternative to static rule-based systems. However, supervised models like XGBoost rely on labeled data, presenting challenges: fraud labeling is costly and requires expert validation, rule-based labels are rigid and potentially omit emerging fraud cases, and supervised models fail to detect novel fraud patterns outside their training data.

To address these limitations, we propose a framework that combines PySpark-based big data processing with hybrid learning using XGBoost for supervised benchmarking and unsupervised methods including Autoencoder (AE), Isolation Forest (IF), and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for anomaly detection. The novel Fraud Relevance Index (FRI) integrates these outputs to prioritize high-risk cases, while SHapley Additive exPlanations (SHAP) analysis provides model explainability for expert validation.

This paper is structured as follows: Section II reviews related work, Section III formulates the problem and describes data processing, Section IV presents the data processing pipeline, Section V presents the proposed framework, Section VI discusses experimental results and Section VII concludes with key findings and future work.

II. Related Work

Telecommunications fraud detection has evolved from rule-based systems to ML-based approaches. With SMS fraud causing global losses of $5.8 billion in 2023 [1], effective detection methods are crucial.

Traditional rule-based systems detect known patterns like SIM boxing through predefined thresholds [2], [3], but struggle with evolving fraud tactics and require frequent manual updates [4], [5]. Supervised ML approaches using XGBoost and ensemble methods show strong performance on labeled data [10], yet remain limited by costly labeling requirements and inability to detect novel patterns. Content-based SMS analysis [6], [7] faces privacy constraints, making CDR-based approaches more practical.

Unsupervised methods address labeling limitations through anomaly detection but often produce high false positive rates requiring refinement. Recent hybrid approaches combining supervised and unsupervised learning show promise [8]–[10], using voting mechanisms and anomaly integration to balance accuracy and adaptability.

Our work extends hybrid approaches by introducing the Fraud Relevance Index for systematic anomaly prioritization and leveraging Big Data processing for scalability, addressing key limitations in existing methods.

III. Problem Formulation

We base our analysis on CDRs to identify fraudulent activities.

A. Data Representation

CDRs contain structured records of telecommunication activities, represented as:

$$X = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^m \tag{1}$$

where:

- X represents the dataset of CDRs.
- Each x_i is a feature vector with m attributes (e.g., pseudonymized sender and receiver Mobile Station International Subscriber Directory Number (MSISDN), timestamp, duration and SMS metadata).
- The dataset is inherently imbalanced, as fraudulent cases are significantly fewer than legitimate ones.

B. Objective Functions

The fraud detection task is formulated as a *binary classification* problem for the *supervised approach* and an *anomaly detection* problem for the *unsupervised approach*.

1) Supervised Approach (XGBoost): Given a labeled dataset (X, Y) where $Y \in \{0, 1\}$ (0: legitimate, 1: fraudulent), we train a classifier:

$$f(X) = Y \tag{2}$$

The model minimizes a *binary cross-entropy loss*:

$$L(y, \hat{y}) = -\sum_{i=1}^{n} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{3}$$

Due to the class imbalance, SMOTE and Tomek Links are applied to improve fraud detection performance, both of which will be explained later in the paper.

2) Unsupervised Approach (AE, IF, DBSCAN): Unsupervised learning is incorporated to detect anomalies and structure fraud cases into meaningful groups.

- Autoencoder: Learns to reconstruct normal SMS behavior. Fraud cases exhibit *higher reconstruction errors*, detected when:

  $$RE(x) = ||x - AE(x)||^2 > \delta \tag{4}$$

  where δ is a predefined threshold [11].
- Isolation Forest: Detects anomalies by computing an anomaly score:

  $$s(x) = 2^{-\frac{E(h(x))}{c(n)}} \tag{5}$$

  where $h(x)$ is the depth of the tree isolating x and $c(n)$ is the average path length of unsuccessful searches in a Binary Search Tree, used to normalize the anomaly score across different sample sizes [10].
- DBSCAN: Groups detected anomalies into fraud clusters and identifies emerging fraud scenarios. Fraud cases are structured based on:
  - Cluster Compactness ($D_{\text{compactness}}$): Measures intra-cluster distances, where tighter clusters suggest structured fraud activity.
  - Cluster Density (D_{density}): Assesses fraud case concentration within clusters, prioritizing highly dense fraud groups.

3) Fraud Relevance Index: Since unsupervised models generate many anomalies, a structured fraud-ranking mechanism is introduced and defined later in section V.

This index ensures that high-risk fraud cases are prioritized while filtering potential false positives.

4) Explainability with SHAP: SHAP is a model-agnostic method that explains how individual features contribute to a model's predictions. It is based on cooperative game theory, where each feature's impact is assessed by measuring how predictions change when the feature is included or excluded in different combinations [12].

Internally, SHAP estimates feature importance by generating perturbed inputs with masked features, observing prediction changes, and assigning fair contribution scores to each feature using model-specific optimizations (e.g., TreeSHAP for XGBoost). By leveraging SHAP, fraud analysts can interpret which features most influence fraud detection decisions, improving transparency and facilitating expert validation.

IV. Data Processing

A. Data Collection and Storage

Given the large volume of CDRs, a Big Data processing pipeline is essential to efficiently handle, store and process the data.

To ensure scalability, Apache Spark (PySpark) is used for distributed data processing, allowing for high-speed transformation of raw data into a structured format suitable for machine learning.

B. Data Preprocessing and Feature Engineering

After collection, the data undergoes multiple preprocessing steps to transform it into a structured format suitable for analysis. These steps include data cleaning, pseudonymization of certain attributes to ensure privacy compliance and feature engineering to extract relevant patterns for fraud detection.

- Data Cleaning: Removal of duplicate records, handling of missing values and filtering out irrelevant attributes to improve data quality.
- Feature Engineering: Extraction of meaningful fraud-related features, including:
  - Volume-based features: Including message frequency per sender, unique recipients, as well as max, mean and variance of sent and received SMS.
  - Temporal patterns: Including night-time SMS ratio, burstiness score, inter-SMS time variance and response time variability.
  - Network-based features: Including direction of communication and reciprocity score (ratio of mutual SMS exchanges to total sent messages, distinguishing one-way fraud from genuine communication).

- Data Aggregation: Processing large-scale CDRs in *PySpark*, followed by aggregation to generate structured datasets compatible with standard ML frameworks.

C. Handling Imbalanced Data

Fraudulent cases are significantly underrepresented in the dataset, making it crucial to address class imbalance. To mitigate this issue in the supervised learning approach, we apply Synthetic Minority Over-sampling Technique (SMOTE) and Tomek Links:

- SMOTE: Generates synthetic fraud samples by interpolating between existing fraud cases, expanding the minority class distribution to improve classification performance [13].
- Tomek Links: Removes overlapping majority class samples that are too close to minority class samples, refining the dataset for better classification [14].

Since unsupervised methods do not rely on labeled data, SMOTE and Tomek Links are not applicable in that context. Instead, anomaly detection techniques such as Autoencoders, Isolation Forest and DBSCAN are used to identify fraud patterns without requiring class labels.

D. Final Dataset Preparation

The preprocessed data is split into training, validation and test sets. The supervised learning model (XGBoost) is trained using the labeled dataset, while the unsupervised models (AE, IF and DBSCAN) process the entire dataset without label dependency. The final dataset structure is optimized to balance efficiency and detection accuracy.

This processed dataset serves as the foundation for the proposed hybrid fraud detection framework, detailed in the next section.

V. Proposed Framework

A. Hybrid Supervised-Unsupervised Learning Pipeline

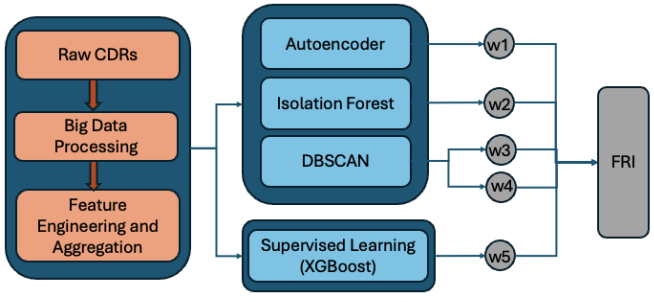

Fig. 1. Hybrid Fraud Detection Pipeline

The fraud detection pipeline follows a structured sequence of anomaly detection, fraud scenario structuring and supervised model benchmarking.

1) Stage 1: Autoencoder-Based Anomaly Detection: An AE is trained on SMS traffic to learn a compressed representation of typical messaging behavior. Given the significant class imbalance in the dataset, the AE primarily learns the dominant normal behavior, as fraudulent cases constitute only a small fraction of the data. As a result, instances that deviate from this learned normal pattern exhibit *higher reconstruction errors*, making them fraud candidates. These errors are then added as a new feature to enrich the dataset.

2) Stage 2: Outlier Detection with Isolation Forest: The dataset, now augmented with reconstruction errors, is passed through Isolation Forest to assign anomaly scores. IF detects fraudulent behavior by evaluating how easily instances are isolated in a randomly partitioned feature space. The higher the anomaly score, the greater the deviation from normal behavior.

3) Stage 3: Clustering with DBSCAN: To structure detected anomalies, we apply DBSCAN. It categorizes anomalies into fraud clusters, grouping cases linked to known fraud scenarios while flagging potential new fraud patterns. The method distinguishes high-density fraud activities from isolated outliers, improving the interpretability of detected fraud cases.

B. Fraud Relevance Index for Fraud Prioritization

Since unsupervised models often generate many anomalies, we introduce the FRI to rank fraud cases based on multiple factors:

$$FRI(x) = \sum_{i=1}^{5} w_i \cdot f_i(x) \tag{6}$$

where the components of $f_i(x)$ are defined as follows:

- $f_1(x) = RE(x)$: Reconstruction error from AE, measuring how much an instance deviates from normal behavior. Higher values indicate greater anomaly likelihood.
- $f_2(x) = S_{IF}(x)$: Anomaly score from IF, assessing how easily an instance is isolated within randomly partitioned decision trees. Higher scores suggest higher anomaly probability.
- $f_3(x) = D_{\text{compactness}}(x)$: Cluster compactness, quantifying the internal consistency of detected fraud clusters. Higher values indicate well-formed fraud patterns rather than scattered anomalies.
- $f_4(x) = D_{\text{density}}(x)$: Cluster density, representing the concentration of fraud cases within a given cluster. Denser clusters are more indicative of systematic fraud.
- $f_5(x) = C_{\text{risk}}(x)$: Fraud risk alignment, measuring the overlap between detected fraud clusters and known fraud cases identified by XGBoost. Higher values reinforce confidence in fraudulent classification.

FRI enables ranking of fraud cases, filtering false positives while prioritizing high-risk anomalies that warrant expert review.

Ideally, the weights w_i in (6) would be optimized via grid search on a labeled validation set. However, due to limited labeled data, we used trial and error guided by domain knowledge to set the weights, which were set as: $w_1 = 0.3$

(reconstruction error), $w_2 = 0.25$ (IF score), $w_3 = 0.2$ (compactness), $w_4 = 0.15$ (density), $w_5 = 0.1$ (risk alignment).

C. Explainability

We integrate SHAP explainability to analyze feature contributions, providing transparency in model decisions and allowing domain experts to evaluate the obtained results.

VI. Experimental Results

This section evaluates the effectiveness of the proposed hybrid fraud detection framework in detecting both known and previously unidentified fraud patterns. Fraud labels were derived from a rule-based approach and validated by experts. While expert validation reduces false positives, rule-based detection relies on hard thresholds, potentially missing some fraud cases.

A. Dataset Characteristics

The experimental evaluation demonstrates the framework's ability to handle real-world data at scale. The original dataset comprised 347 million CDRs collected over one month, containing 182 raw attributes per record across 29,567 unique MSISDNs. This massive volume of granular transaction data required sophisticated distributed processing to extract meaningful fraud patterns.

Using PySpark's distributed computing capabilities, we performed extensive data preprocessing and feature engineering on this large-scale dataset. The process involved filtering SMS-specific transactions, computing complex temporal and behavioral aggregations per MSISDN (such as SMS frequency ratios, burst patterns, reciprocity scores, and cross-network communication patterns), and applying domain expertise to identify the most discriminative fraud indicators. Feature selection reduced raw CDR attributes to 67 fraud-relevant features.

The significant data reduction enables the application of standard machine learning libraries while preserving the essential behavioral signatures necessary for accurate fraud detection. The resulting dataset structure allows for comprehensive model evaluation while maintaining computational efficiency and interpretability.

B. Evaluation of Supervised Learning (XGBoost)

XGBoost serves as a benchmark for detecting known fraud cases. Its performance is shown in Figure 2.

XGBoost achieves high precision (0.92) and recall (0.83) but is limited by rule-based labeling, which may fail to capture emerging fraud patterns. This highlights the need for unsupervised learning to detect fraud beyond predefined rules.

C. Evaluation of Autoencoder-Based Anomaly Detection

The Autoencoder, learns normal behavior of SMS behavior. Fraudulent messages exhibit higher reconstruction errors, as shown in Figure 3.

AE identifies anomalies through reconstruction error but produces continuous scores rather than binary fraud classifications. These scores are combined with IF and DBSCAN outputs via the Fraud Relevance Index to create a weighted composite score that eliminates the need for manually setting individual thresholds for each unsupervised method.

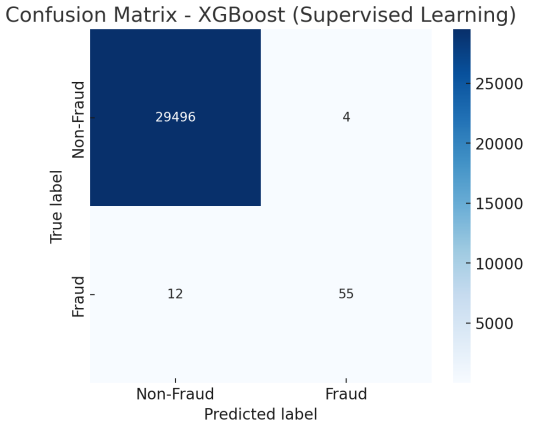

Fig. 2. Confusion Matrix of XGBoost (Supervised Learning)

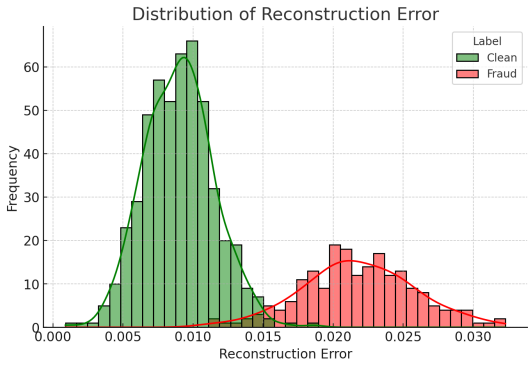

Fig. 3. Normal vs Fraudulent Reconstruction Errors

D. Evaluation of Hybrid Approach

The FRI integrates anomaly scores, clustering and supervised fraud labels to improve fraud detection. Its performance is shown in Figure 4.

Compared to XGBoost, FRI:

- Achieves higher recall (0.88), detecting more fraud cases while maintaining competitive precision.
- Demonstrates fewer false positives than pure unsupervised learning approaches, benefiting from the integration of supervised model insights and clustering refinement.
- Successfully identified fraud cases belonging to categories not represented in the original rule-based training set, demonstrating the framework's ability to detect emerging or previously unclassified fraud patterns that fall outside the scope of existing detection rules.

Due to time constraints, only a subset of flagged fraud cases was manually reviewed. Further validation may reveal additional cases missed by the rule-based approach.

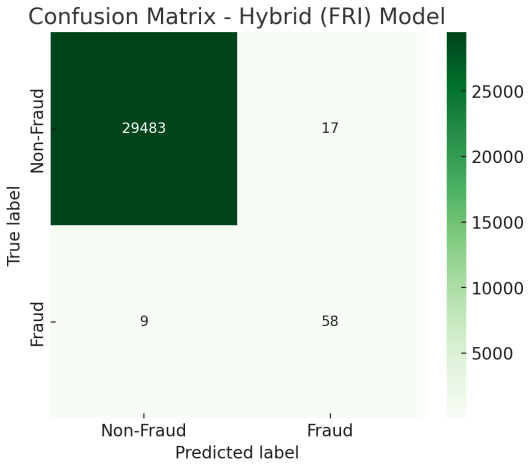

Fig. 4. Confusion Matrix of Hybrid Approach (FRI)

E. SHAP Feature Importance for Interpretability

To improve interpretability, SHAP values highlight key fraud indicators (Figure 5).

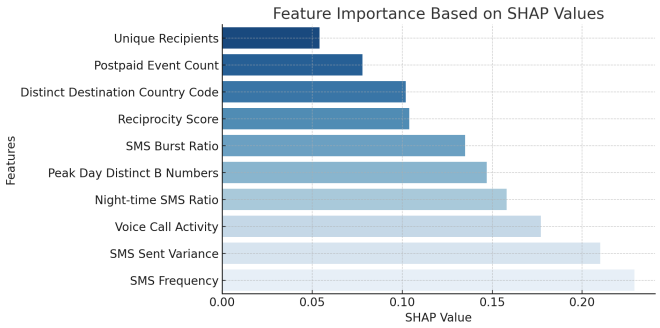

Fig. 5. Feature Importance Based on SHAP Values

F. Discussion

The hybrid approach improves fraud detection by detecting more fraud cases beyond rule-based labels, processing large-scale datasets efficiently with PySpark, and enhancing interpretability through SHAP analysis for expert validation.

Manual validation of more fraud cases is necessary to further assess FRI's real-world effectiveness. The confirmed detection of an undetected fraud case highlights FRI's potential to identify evolving fraud patterns, making it a valuable tool for adaptive fraud detection.

VII. Conclusion and Future Work

A. Conclusion

This study proposed a hybrid machine learning framework for telecommunication fraud detection, leveraging Big Data processing with PySpark to efficiently handle large-scale CDRs.

The framework integrates supervised and unsupervised learning to overcome rule-based system limitations. The Fraud Relevance Index prioritizes fraud cases by combining anomaly scores, clustering properties and supervised model alignment. SHAP-based explainability enhances transparency, aiding expert validation.

Experimental results demonstrate improved fraud detection accuracy, reduced false positives and adaptability to evolving fraud tactics.

B. Future Work

Although our framework achieves promising results, several areas remain open for further research and improvement: implementing real-time streaming detection to identify fraud as it occurs, expanding the framework to detect broader telecommunication fraud categories beyond SMS fraud, and integrating semi-supervised learning where newly detected fraud patterns can iteratively refine the supervised model.

These future enhancements will further improve the adaptability, efficiency and robustness of fraud detection in telecommunications, ensuring better fraud prevention strategies in evolving environments.

References

[1] Communications Fraud Control Association, "Global Fraud Loss Survey," 2023.
[2] B. Abo Yehya et al., "Telecommunications Fraud: Machine Learning-Based Detection," *Int. J. Data Sci.*, 2021.
[3] P. Ghosh and S.-H. Kim, "Telecom Fraud Detection with Machine Learning on Imbalanced Dataset," in *Proc. IEEE Conf. Mach. Learn. Appl.*, 2020.
[4] R. Li, Y. Zhang, Y. Tuo and P. Chang, "A Novel Method for Detecting Telecom Fraud Users," in *Proc. Int. Conf. Inf. Syst. Eng. (ICISE)*, 2018, pp. 46–50.
[5] L. Schmidt and T. Becker, "Improving a Rule-Based Fraud Detection System with Classification Based on Association Rule Mining," in *Proc. Eur. Conf. Artif. Intell. (ECAI)*, 2022.
[6] S. R. Prusty, B. Sainath, S. K. Jayasingh and J. K. Mantri, "SMS Fraud Detection Using Machine Learning," in *Proc. IEEE Conf. Cyber Secur.*, 2021.
[7] A. Taylor and A. Robert, "Using Machine Learning to Detect Fraudulent SMSs in Chichewa," in *Proc. Int. Conf. Mach. Learn.*, 2022.
[8] S. Daskalaki, I. Kopanas, M. Goudara and N. Avouris, "Data Mining for Decision Support on Customer Insolvency in Telecommunications," *Eur. J. Oper. Res.*, vol. 145, no. 2, pp. 239–255, 2003.
[9] A. Khan and J. Wang, "Machine Learning-Based Fraudulent Detection System for Financial Transactions," *IEEE Trans. Neural Netw. Learn. Syst.*, 2021.
[10] F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," 2008 Eighth *IEEE International Conference on Data Mining, Pisa, Italy*, 2008.
[11] P. Vincent, H. Larochelle, Y. Bengio and P. A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. 25th Int. Conf. Mach. Learn.*, Helsinki, Finland, 2008, pp. 1096–1103.
[12] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 4768–4777.
[13] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "SMOTE: synthetic minority oversampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.
[14] I. Tomek, "Two modifications of CNN," *IEEE Trans. Syst. Man Cybern.*, vol. SMC-6, no. 11, pp. 769–772, Nov. 1976.