Boosting Fraud Detection in Vietnam's Electricity Systems with a Sampling-Based Imbalance Handling Module in Ensemble Learning Models

Phung Do[1], Dinh-Phat Nguyen[1], Van-Hien Nguyen[1], Thang Cap[1], and Tuong Le[2,*]

[1] University of Information Technology, Vietnam National University Ho Chi Minh City, Vietnam
phungdtm@uit.edu.vn, thangcpd@uit.edu.vn, 23521144@gm.uit.edu.vn, 21522064@gm.uit.edu.vn

[2] Faculty of Information Technology, HUTECH University, Ho Chi Minh City, Vietnam
lc.tuong@hutech.edu.vn

Abstract. Electricity fraud detection remains a critical challenge for utility providers, especially in developing countries where data is large-scale and highly imbalanced. This paper introduces an adaptive resampling framework called the Hybrid Sampling Ratio Selection Algorithm (HSRSA), designed to enhance fraud detection performance in smart grid systems. The method systematically explores combinations of random undersampling and SMOTE-based oversampling ratios to optimize classifier sensitivity under low false positive rate (FPR) constraints. We evaluate the effectiveness of HSRSA across four ensemble-based models—CatBoost, LightGBM, XGBoost, and Decision Tree—using a real-world dataset from Vietnam Electricity (EVN) comprising over 22 million smart meter records. Experimental results demonstrate that the proposed approach significantly improves partial AUC (pAUC) scores compared to baseline resampling, particularly in the critical FPR <= 0.2 region. The CatBoost model, when combined with the optimal sampling ratio identified by HSRSA, achieves a pAUC of 0.1992, outperforming its baseline counterpart. This study highlights the importance of ratio tuning in hybrid resampling and contributes a scalable, model-agnostic solution for electricity theft detection in imbalanced and large-scale settings.

Keywords: Fraud Detection, Ensemble Learning · CatBoost · LightGBM · Hybrid Sampling

1 Introduction

Electricity fraud poses a major threat to the sustainability and financial integrity of power distribution systems, particularly in developing countries like Vietnam. These fraudulent activities—such as meter tampering, illegal connections, or manipulation of digital meters—lead to substantial losses for utility companies and disrupt the stability of national energy systems. According to EVN (Vietnam Electricity), cities like Ho Chi Minh City alone lose over 18 million kWh annually due to theft, equating to approximately 32 billion VND. Traditional detection methods, which rely heavily on manual inspection, are not only labor-intensive but also inadequate in capturing the evolving complexity of fraud behaviors.

The emergence of machine learning (ML) and ensemble learning methods has introduced new avenues for automating fraud detection. Ensemble classifiers such as Decision Trees, XGBoost, LightGBM, and CatBoost have demonstrated significant effectiveness in handling high-dimensional, large-scale datasets like those derived from smart meters [1]. Decision Trees, as a foundation, offer intuitive rule-based structures for classification but tend to overfit without pruning or ensemble enhancements [1]. XGBoost, an advanced gradient boosting algorithm, introduces regularization and parallel computation to overcome Decision Tree limitations and has become widely used in fraud detection challenges due to its robustness and scalability [2]. Similarly, LightGBM improves training efficiency by using histogram-based learning and leaf-wise tree growth [3]. CatBoost further optimizes boosting by effectively processing categorical variables while addressing prediction shift and target leakage, making it suitable for imbalanced tabular datasets [3].

However, these ensemble models face difficulties when applied to imbalanced datasets, where fraudulent transactions constitute only a small portion of total records. This skew in class distribution can cause models to favor the majority class, diminishing their effectiveness in detecting rare but critical fraud instances. To mitigate this, researchers have proposed hybrid sampling strategies, including random under-sampling and SMOTE-based over-sampling, to rebalance training data [4]. A study by Almubark et al. [4] emphasized that combining ensemble learning with adaptive sampling and thresholding techniques can significantly enhance precision and recall in highly imbalanced settings.

Building upon this foundation, the present study introduces a novel Hybrid Sampling Ratio Selection Algorithm to optimize the resampling process for ensemble classifiers. By dynamically identifying the most effective combination of under- and over-sampling ratios, the proposed method aims to maximize detection performance—especially under low false positive constraints—measured through Area Under the Curve (AUC) and partial AUC (pAUC). This research contributes a scalable and adaptive approach tailored for real-world electricity fraud detection in Vietnam's evolving energy landscape.

2 Related Work

Fraud detection has long been a critical area of research in domains ranging from finance to energy systems. This section categorizes and synthesizes prior studies into three themes: general fraud detection methods, electricity theft detection techniques, and advanced learning approaches for large-scale imbalanced datasets.

In the context of real-time credit card fraud detection, Gupta et al. [5] demonstrated that incorporating temporal behavioral features and regional transaction analysis can significantly enhance detection rates. Moreover, they emphasized the use of AUC and partial AUC (pAUC) as robust metrics beyond simple accuracy.

Pamir et al. [6] proposed a two-stage thresholding ensemble method combined with random undersampling. Their model achieved improved recall and reduced false positives, which are critical in practical fraud detection settings. Similarly, Iftikhar et al. [7] introduced deep attention-based architectures tailored for big imbalanced data, showing improved generalisation and fraud classification accuracy. Electricity theft detection has emerged as a crucial research area due to its direct economic impact on utilities and national grid operations. Mbey et al. [8] introduced a hybrid deep learning model combining BiLSTM with convolutional attention mechanisms to capture both temporal and spatial patterns in electricity consumption. This approach outperformed traditional methods under non-stationary data conditions.

Almalki et al. [9] proposed an ensemble framework that leverages contextual feature selection and temporal behavioral profiling. Their results showed superior performance in both precision and recall, particularly when using boosting-based models such as XGBoost and CatBoost in smart metering environments. To address issues of incomplete or noisy data, Theodorakopoulos et al. [10] developed an unsupervised label generation pipeline based on gradient boosting models. The approach not only improved classification performance but also provided interpretability via explainable features.

Recent studies have focused on scalable learning solutions tailored to real-world fraud detection, which often involves high-volume and highly imbalanced datasets. Walauskis et al. [11] provided a comprehensive survey of deep learning models, underscoring the importance of adaptive architectures capable of handling evolving fraudulent behavior (i.e., concept drift). In [12], Aburbeian et al. proposed a multimodal detection framework using categorical embeddings and gradient boosting methods such as XGBoost and CatBoost, demonstrating robustness on high-dimensional electricity datasets. Zhu et al. [13] explored federated learning to address data privacy and decentralization challenges in energy fraud detection. Their model showed scalability and effectiveness in distributed environments.

Finally, Hasnony et al. [14] applied SMOTE-based oversampling and ensemble classifiers to large telecom and energy datasets. Their results highlighted that properly tuned resampling strategies can significantly enhance recall for fraudulent instances without sacrificing precision.

3 Dataset and Methods

3.1 Dataset

The dataset comprises 22,178,305 observations spanning 62 distinct features, collected from multiple provinces in Vietnam. The classification target variable, referred to as Electricity Fraud, is encoded as a binary outcome: fraudulent (1) and non-fraudulent (0). To ensure robust model development, hyperparameter tuning, and unbiased performance evaluation, the dataset was systematically divided into two subsets: training and testing. Specifically, the training set includes 17,517,989 non-fraudulent and 224,655 fraudulent records. The testing set comprises 4,379,498 non-fraudulent and 56,163 fraudulent instances, facilitating independent evaluation of the model's generalization capability.

Given the large number of features and the presence of redundant or less-informative variables, a feature selection strategy was adopted to improve model efficiency and reduce the risk of overfitting. Instead of utilizing all 61 available input features (excluding the target variable), the proposed model employs a carefully curated subset of 27 features. These selected features were chosen based on their domain relevance and predictive importance, ensuring that the model captures the most informative aspects of the data. The selected features can be broadly categorized into four groups: (i) technical attributes, such as voltage levels, current readings, and energy index values; (ii) temporal dynamics, including metering start and end timestamps as well as segmented energy usage patterns; (iii) customer-specific information, including prior and current meter readings along with derived consumption values; and (iv) geospatial and device metadata, such as device codes, customer location, and regional identifiers. This structured feature set enhances the model's ability to detect abnormal consumption behaviors and fraudulent activities with greater precision and robustness. This structured feature subset ensures comprehensive representation of consumption behavior, operational context, and user metadata—enabling the learning models to more effectively detect anomalous usage patterns indicative of fraudulent activities.

To illustrate the pronounced class imbalance present in the dataset partitions, Figure 1 displays the distribution of fraudulent and non-fraudulent records across the training, testing subsets.

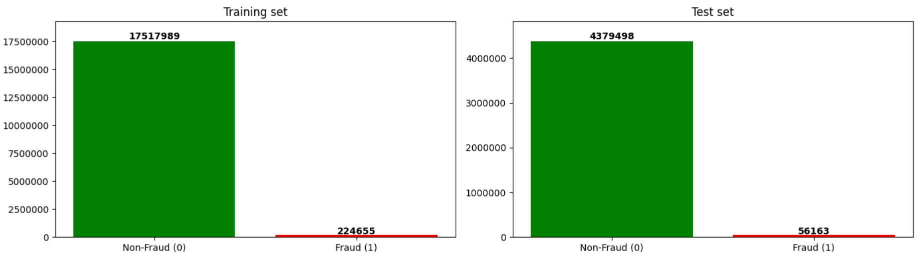

Fig. 1. Comparison of Fraudulent and Non-Fraudulent in Electricity Data.

As depicted in Figure 1, fraudulent cases (shown in red) constitute only a small fraction relative to the overwhelming number of non-fraudulent records (shown in green) in every subset. This severe imbalance highlights the necessity for advanced imbalance handling strategies, as standard machine learning models may otherwise be biased towards the majority class. Addressing this issue is critical to ensure effective and reliable fraud detection, as discussed in the following sections.

3.2 System architecture

The overall system architecture of the proposed fraud detection framework is illustrated in Figure 2. It comprises two major modules: (1) data preparation and (2) hybrid sampling ratio selection algorithm. This architecture is designed to systematically address data imbalance while optimizing classifier performance on real-world electricity consumption data.

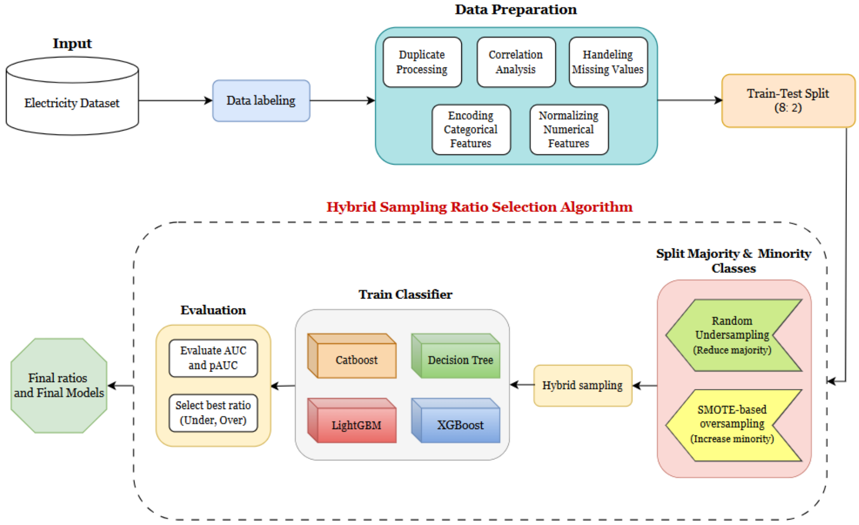
Input
Electricity Dataset
Data labeling
Data Preparation
Duplicate Processing
Correlation Analysis
Handeling Missing Values
Encoding Categorical Features
Normalizing Numerical Features
Train-Test Split (8: 2)
Hybrid Sampling Ratio Selection Algorithm
Split Majority & Minority Classes
Random Undersampling (Reduce majority)
SMOTE-based oversampling (Increase minority)
Hybrid sampling
Train Classifier
Catboost
Decision Tree
LightGBM
XGBoost
Evaluation
Evaluate AUC and pAUC
Select best ratio (Under, Over)
Final ratios and Final Models

Fig. 2. The system Architecture of Fraud Detection using a Hybrid Sampling Ratio Selection Algorithm.

Data Preparation Module

This module begins with the import of raw electricity consumption data from the EVN dataset. The initial stage involves data labeling, where each record is tagged as either fraudulent (1) or non-fraudulent (0). Following this, a comprehensive data preprocessing pipeline is applied, including:

- **Duplicate Processing:** All records were scanned for duplicate entries. Any duplicate rows—identified based on identical feature values across all columns—were removed to avoid redundancy and potential bias in the learning process.

- **Correlation Analysis:** Feature correlation analysis was performed to identify and address multicollinearity among input variables. Highly correlated features were either merged, removed, or otherwise adjusted to enhance model interpretability and reduce overfitting.
- **Handling Missing Values:** Missing or incomplete data points were systematically processed. Numerical features with missing values were imputed using the median value of each feature, minimizing the influence of outliers. For categorical variables, missing entries were filled with the mode (most frequent) value. Records with missing values in critical target or identifier fields were discarded.
- **Encoding Categorical Features:** Categorical attributes were encoded to enable machine learning models to process them effectively. Ordinal Encoding was applied to transform categorical variables into integer representations, preserving any inherent order and ensuring compatibility with ensemble learning algorithms.
- **Normalizing Numerical Features:** To standardize the scales of numerical features and avoid dominance of features with larger magnitudes, all continuous variables were normalized using z-score normalization (StandardScaler), resulting in zero mean and unit variance. This step is especially important for algorithms sensitive to feature scale, such as gradient-boosted decision trees.

These preprocessing steps collectively ensured a clean, consistent, and high-quality dataset, forming a robust foundation for subsequent training and evaluation of fraud detection models.

Hybrid Sampling Ratio Selection Module

To address the issue of class imbalance, we propose a hybrid sampling strategy that combines random undersampling of the majority class and SMOTE-based oversampling of the minority class. The key objective is to identify an optimal combination of under- and over-sampling ratios that improves the model's ability to detect fraudulent cases, particularly under low false positive rates.

- **Problem Definition**

  Let the original imbalanced dataset be defined as:

  $$D = \{(x_i, y_i)\}_{i=1}^{n}, \quad y_i \in \{0, 1\}$$

  where $y_i = 0$ represents a non-fraudulent instance and $y_i = 1$ denotes a fraudulent case.

  To address class imbalance, we define two sets of sampling ratios:

  - The set of candidate undersampling ratios:

    $$R_u = \{r_u^1, r_u^2, ..., r_u^m\}$$

  - The set of candidate oversampling ratios:

    $$R_o = \{r_o^1, r_o^2, ..., r_o^n\}$$

The objective is to determine the optimal pair (r_u^*, r_o^*) from the cross-product $R_u \times R_o$ that maximizes the partial Area Under the Curve (pAUC) over a specified False Positive Rate (FPR) range $[0, \alpha]$, where α denotes the maximum acceptable FPR. The optimization is formulated as:

$$(r_u^*, r_o^*) = \underset{r_u \in R_u, r_o \in R_o}{\arg\max} \ \mathrm{pAUC}_{[0,\alpha]}(r_u, r_o)$$

Where:

- (r_u^*, r_o^*): The optimal pair of undersampling and oversampling ratios.
- R_u: The set of candidate undersampling ratios.
- R_o: The set of candidate oversampling ratios.
- $\mathrm{pAUC}_{[0,\alpha]}$: The partial Area Under the ROC Curve, restricted to the interval $[0, \alpha]$ on the FPR axis.
- α: The maximum acceptable false positive rate.
- $\arg\max$: Returns the argument (input values) that maximizes the given function.

This formulation enables the selection of optimal sampling ratios that enhance detection performance, particularly under constraints of low false positive rates, which is crucial in high-risk domains such as fraud detection.

Algorithm 1: Hybrid Sampling Ratio Selection Algorithm

```
Input:
  • Imbalanced dataset $\mathcal{D}$
  • Candidate undersampling ratios $R_u$ and oversampling ratios $R_o$
  • Maximum allowed false positive threshold $\alpha$
Output: Optimal hybrid sampling ratio $(r_u^*, r_o^*)$
Initialize $best\_score \leftarrow 0$ ;
Initialize $best\_ratio \leftarrow (0, 0)$ ;
foreach $(r_u, r_o) \in R_u \times R_o$ do
    Apply random undersampling with ratio $r_u$ to the majority class;
    Apply SMOTE oversampling with ratio $r_o$ to the minority class;
    Train classifier $M$ (e.g., LightGBM) on the resampled dataset;
    Evaluate the model on the test set using partial AUC:;
        $pAUC_{[0,\alpha]}(r_u, r_o) = \int_0^{\alpha} TPR(FPR)\, dFPR$;
    if $pAUC_{[0,\alpha]}(r_u, r_o) > best\_score$ then
        $best\_score \leftarrow pAUC_{[0,\alpha]}(r_u, r_o)$;
        $best\_ratio \leftarrow (r_u, r_o)$;
return $best\_ratio$
```

This study presents a new method to fix the problem of class imbalance in electricity fraud detection: the Hybrid Sampling Ratio Selection Algorithm. This method looks at different combinations of undersampling and oversampling ratios to rebalance the dataset before training. It checks each setup using the partial Area Under the Curve (pAUC) in a low false positive rate (FPR) area. This way, the algorithm finds the best sampling ratio pair to improve detection performance. This approach boosts the ability of ensemble classifiers to identify fraud while also reducing false alarms.

4 Experiments

4.1 Experimental Setting

All experiments were performed on a high-performance workstation equipped with an AMD64 multi-core CPU, 112 GB of RAM, and running the 64-bit Microsoft Windows 11 OS. This computational environment was optimized for efficient processing of large-scale datasets, ensuring the reproducibility of all experimental protocols. In this case study, we assessed the performance of four tree-based classifiers: LightGBM, XGBoost, CatBoost, and Decision Tree. Each model was meticulously configured to facilitate a fair and reproducible comparison.

The experiments employed four tree-based classifiers, each configured with specific hyperparameters to balance model complexity and performance:

- **LightGBM**: Implemented using the gradient boosting decision tree framework with `boosting_type` set to `gbdt`. The model was trained for 100 boosting iterations (`n_estimators` = 100) with a learning rate of 0.1. The maximum number of leaves was limited to 31 (`num_leaves` = 31), and a minimum of 3 samples was required per leaf node (`min_child_samples` = 3). The minimum child weight was set to 0.001 (`min_child_weight` = 0.001). To ensure reproducibility, a fixed random seed (`random_state` = 42) was used, and output verbosity was suppressed (`verbosity` = -1).
- **XGBoost**: Configured with 100 boosting iterations (`n_estimators` = 100) and a maximum tree depth of 4 (`max_depth` = 4) to control overfitting while capturing important feature interactions. A minimum child weight of 3 (`min_child_weight` = 3) was set for regularization. The histogram-based algorithm (`tree_method` = `hist`) was utilized for efficient training and reduced memory usage. The random seed was set to 42 (`random_state` = 42).
- **CatBoost**: Trained for 100 boosting iterations (`iterations` = 100) with a learning rate of 0.1. The maximum depth was set to 16 (`depth` = 16), and each leaf node required a minimum of 3 samples (`min_data_in_leaf` = 3). For consistency, the random seed was set to 42 (`random_state` = 42).
- **Decision Tree**: Configured with a maximum depth of 6 (`max_depth` = 6) to prevent overfitting. The Gini impurity criterion (`criterion` = `gini`) was used for evaluating splits. Internal nodes required at least 10 samples to split (`min_samples_split` = 10), and terminal leaf nodes needed a minimum of 5 samples (`min_samples_leaf` = 5). The splitter was set to `best`, and a fixed random seed (`random_state` = 42) ensured reproducibility.

All hyperparameters were selected based on recommended defaults or preliminary tuning to balance computational efficiency and model generalization.

4.2 Evaluation Metrics

To rigorously evaluate the performance of the proposed ensemble-based framework for electricity fraud detection, this study employs two principal metrics: the **Area Under the Receiver Operating Characteristic Curve (AUC)** and the **Partial AUC (pAUC)**.
The **ROC curve** is a graphical plot that characterizes the trade-off between the true positive rate (TPR) and the false positive rate (FPR) of a binary classifier at various threshold settings. Specifically:

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN} \tag{1}$$

where TP, FP, TN, and FN denote the numbers of true positives, false positives, true negatives, and false negatives, respectively.
The AUC is defined as the area under the ROC curve, quantifying the model's overall discriminatory power:

$$AUC = \int_0^1 TPR(FPR)\, dFPR \tag{2}$$

The **pAUC** measures the area under the ROC curve within a critical operational range of FPR, reflecting the model's performance in minimizing false positives:

$$pAUC_{[0,\alpha]} = \int_0^\alpha TPR(FPR)\, dFPR \tag{3}$$

where:
- $TPR(FPR)$ is the true positive rate as a function of the false positive rate (FPR),
- $dFPR$ denotes integration with respect to the FPR, computing the area under the TPR curve over the FPR range,
- α is the maximum FPR threshold of interest (e.g., $\alpha = 0.2$ or another operationally significant value).

The pAUC measures the normalized area under the ROC curve from FPR $= 0$ to FPR $= \alpha$, providing a focused assessment of model performance in the low-FPR region, which is particularly relevant for applications such as fraud detection.

4.3 Experimental Results

Table 1 provides a comprehensive comparison of the performance of four tree-based classifiers—CatBoost, Decision Tree, XGBoost, and LightGBM—with and without the application of the Hybrid Sampling Ratio Selection Algorithm. The evaluation focuses on two key metrics: the Area Under the ROC Curve (AUC), which measures overall discriminative ability, and the partial AUC (pAUC), which specifically captures model performance within the operationally critical region of low false positive rates.

Table 1. Performance comparison with and without using Hybrid Sampling Ratio Selection Algorithm

No	Models	AUC		pAUC	
		Baseline	**Hybrid Sampling**	**Baseline**	**Hybrid Sampling**
1	CatBoost	0.9988	0.9990	0.1991	0.1992
2	Decision Tree	0.9285	0.9629	0.1425	0.1835
3	XGBoost	0.9231	0.9964	0.1292	0.1981
4	LightGBM	0.9933	0.9973	0.1944	0.1984

Baseline(Imbalaced Dataset)

When trained on the original, imbalanced dataset, CatBoost and LightGBM both demonstrated strong overall performance, with AUCs of 0.9988 and 0.9933, and pAUCs of 0.1991 and 0.1944, respectively. In contrast, the Decision Tree and XGBoost models exhibited substantially lower scores, with Decision Tree achieving an AUC of 0.9285 and pAUC of 0.1425, and XGBoost recording an AUC of 0.9231 and pAUC of 0.1292. These findings highlight the sensitivity of conventional decision tree methods to class imbalance, particularly in the context of operationally relevant detection thresholds.

With Hybrid Sampling Ratio Selection Algorithm

The application of the Hybrid Sampling Ratio Selection Algorithm led to notable improvements across all models, most dramatically for those initially more affected by imbalance. Decision Tree's AUC rose from 0.9285 to 0.9629 and its pAUC from 0.1425 to 0.1835, while XGBoost improved from an AUC of 0.9231 to 0.9964 and a pAUC of 0.1292 to 0.1981. These increases reflect a significantly enhanced ability to accurately identify fraudulent instances while maintaining a low false positive rate, an essential requirement for real-world deployment in electricity fraud detection.

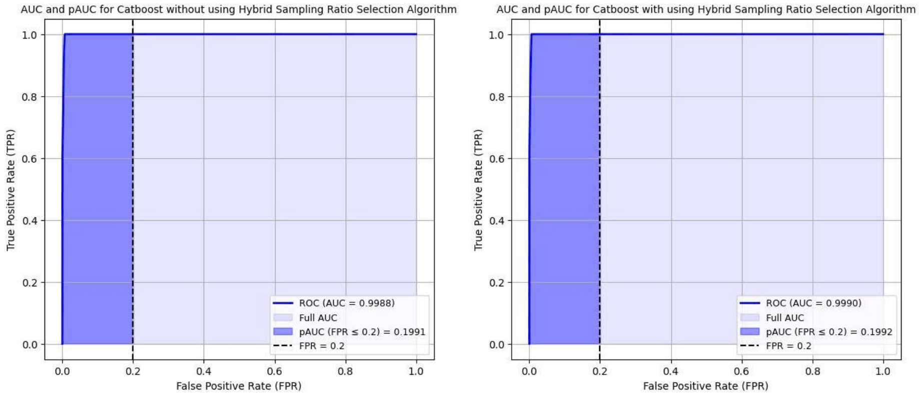

Fig. 3. AUC and pAUC for Catboost model.

CatBoost and LightGBM, which already exhibited robust performance in the baseline scenario, also benefited from the hybrid sampling strategy. The performance comparison of the CatBoost model with and without the proposed Hybrid Sampling Ratio Selection Algorithm is illustrated in Figure 3. CatBoost's AUC improved from 0.9988 to 0.9990 and its pAUC from 0.1991 to 0.1992. Similarly, LightGBM's AUC increased from 0.9933 to 0.9973 and its pAUC from 0.1944 to 0.1984. While the relative gains for these models are smaller, the results affirm that hybrid sampling does not negatively impact—and may further solidify—the performance of already strong classifiers.

5 Conclusion

This study presents a novel ensemble learning framework for electricity fraud detection in Vietnam, integrating a Hybrid Sampling Ratio Selection Algorithm to address the significant class imbalance inherent in real-world datasets. By combining Random Undersampling and SMOTE-based Oversampling in a grid search strategy, the proposed method identifies optimal resampling ratios that improve model performance—particularly in the low false positive rate region, which is critical for practical applications.

Experimental results demonstrate that the hybrid sampling strategy consistently enhances classification performance across all evaluated models, including CatBoost, LightGBM, XGBoost, and Decision Tree. Notably, substantial improvements in both AUC and pAUC were observed for models traditionally more sensitive to class imbalance, such as XGBoost and Decision Tree. Even high-performing models like CatBoost and LightGBM exhibited measurable gains, confirming the method's generalizability and robustness.

These findings underscore the effectiveness of ratio-tuned hybrid sampling in improving fraud detection accuracy without compromising operational requirements. The algorithm's ability to optimize both global and localized metrics makes it highly suitable for deployment in large-scale fraud monitoring systems.

Building upon the current results, future research directions include:

- Adaptive Ratio Optimization: Incorporating Bayesian optimization or reinforcement learning to dynamically adjust sampling ratios during training, instead of relying on grid search.
- Model-Agnostic Integration: Extending the hybrid sampling module for integration with deep learning classifiers and time-series anomaly detection frameworks.
- Real-time Deployment: Developing an online version of the algorithm for continuous fraud detection in streaming electricity usage data.
- Cross-domain Validation: Applying the framework to other domains such as banking, telecommunications, or insurance, to validate its versatility under different data characteristics.

The proposed Hybrid Sampling Ratio Selection Algorithm provides a strong foundation for intelligent imbalance handling, contributing toward more accurate, operationally efficient, and deployable fraud detection systems in critical infrastructure environments.

References

1. S. R. Safavian, D. Landgrebe: A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics* **21**(3), (1991)
2. T. Chen, C. Guestrin: XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016)
3. L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, A. Gulin: CatBoost: Unbiased Boosting with Categorical Features. *Advances in Neural Information Processing Systems* **31**, 6638–6648 (2018)
4. Almubark, Ibrahim: Advanced Credit Card Fraud Detection: An Ensemble Learning Using Random Under Sampling and Two-Stage Thresholding. *IEEE Access*, (2024)
5. P. Gupta, A. Varshney, M. Rafeek Khan, R. Ahmed, M Shuaib, S. Alam: Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques. *International Conference on Machine Learning and Data Engineering* **218**, 2575-2584 (2023)
6. Pamir Javaid, Nadeem Javed, Muhammad Umar, Mohamad Abou, Abdullah M. and Imran, Muhammad: Electricity theft detection for energy optimization using deep learning models. *Energy Science & Engineering* **11**(10), 3575-3596 (2023)
7. Iftikhar, Hasnain Khan, Nitasha Raza, M. Amir, Ghulam Khan, Murad Aoudia, Mouloud Touti, Ezzeddine Emara: Electricity theft detection in smart grid using machine learning. *Frontiers in Energy Research* **12**, (2024)
8. Mbey, Camille Franklin, Jacques and Yem Souhe, Felix Ghislain, Vinny Junior, Alexandre Teplaira: Electricity Theft Detection in a Smart Grid Using Hybrid Deep Learning-Based Data Analysis Technique. *Journal of Electrical and Computer Engineering* **2024**, 16, (2024)
9. Fahad Almalki, Mehedi Masud: Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods. In: *arXiv preprint arXiv:2505.10050*, Article 10050 (2025)
10. Theodorakopoulos, L.Theodoropoulou, A. Tsimakis, A. Halkiopoulos, Constantinos: Big Data-Driven Distributed Machine Learning for Scalable Credit Card Fraud Detection Using PySpark, XGBoost, and CatBoost. *Electronics* **14** (2025)
11. Walauskis, Mary Anne, Khoshgoftaar, Taghi M.: Unsupervised label generation for severely imbalanced fraud data. *Journal of Big Data* **12**, (2025)
12. Aburbeian, AlsharifHasan Mohamad, Ashqar, Huthaifa I.: Credit Card Fraud Detection Using Enhanced Random Forest Classifier for Imbalanced Data. In: *Proceedings of the 2023 International Conference on Advances in Computing Research (ACR'23)*, 605–616 (2023)
13. M. Zhu, Ye Zhang, Yulu Gong, Changxin Xu, Yafei Xiang: Enhancing Credit Card Fraud Detection A Neural Network and SMOTE Integrated Approach. In: *Computational Engineering, Finance, and Science*, (2024)

14. hasnony, ibrahim, Elfetouh, Ahmed, Rezk, Amira: Enhancing Fraud Detection in Imbalanced Datasets: A Comparative Study of Machine Learning and Deep Learning Algorithms with SMOTE Preprocessing. In: *Mansoura Journal for Computer and Information Sciences* **20** , 1–21 (2025)