Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 3, March : 2025

CREDIT CARD FRAUD DETECTION USING LOGISTIC REGRESSION AND ADAPTIVE TRAINING STRATEGIES

Dr. M. Babu Rao Professor & HOD, Department of CSE, Seshadrirao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh baburaompd@gmail.com
B. Lalitha, B. Dhatrisri, A. Rajesh, A.Kalam, UG Student, Department of CSE, Seshadrirao Gudlavalleru Engineering College, Gudlavalleru, Andhra Pradesh.

Abstract
The increasing reliance on digital transactions has led to a surge in credit card fraud, which poses significant financial risks to individuals and institutions. Traditional fraud-detection systems often struggle because of the highly imbalanced nature of transaction datasets, in which fraudulent cases represent only a small fraction of all transactions. This study addresses these challenges by developing a machine-learning-based fraud detection system using Logistic Regression and adaptive training strategies. To improve the detection accuracy, the Synthetic Minority Oversampling Technique (SMOTE) was applied to handle class imbalance by generating synthetic fraudulent samples. Additionally, feature scaling using StandardScaler ensures that all features are normalized for better model performance. The proposed system was trained and evaluated using various performance metrics, including the accuracy (96.49%), precision (98%), recall (95%), and F1-score (96%). The model achieved an impressive ROC-AUC score of 0.9935, demonstrating its ability to effectively distinguish between legitimate and fraudulent transactions. The results highlight the effectiveness of combining Logistic Regression with data preprocessing techniques to enhance fraud detection in imbalanced datasets.
Keywords: Credit Card Fraud Detection, Logistic Regression, SMOTE, Imbalanced Data, ROC-AUC Score

1. Introduction

In today's digital world, credit cards have become essential for financial transactions, offering convenience and security[1]. However, with the rise in online and cashless payments, credit card fraud has increased, leading to billions of global financial losses[2]. Fraudsters have continuously developed new techniques to bypass security measures, making fraud detection a crucial challenge for banks and financial institutions[3]. Detecting fraudulent transactions efficiently can help prevent financial loss and maintain trust in digital payment systems. One of the biggest challenges in fraud detection is the highly imbalanced nature of the transaction datasets. Because fraudulent transactions account for a tiny fraction of all transactions, traditional machine learning models tend to favor legitimate transactions, often misclassifying fraud as a regular activity[4]. This leads to high false-negative rates where fraudulent transactions are undetected. Another major challenge is the evolving nature of fraudulent techniques[5]. Fraudsters frequently change their tactics, which makes it difficult for static detection systems to adapt. Additionally, real-time fraud detection is critical, as fraud prevention systems must instantly flag and stop unauthorized transactions[6]To address these challenges, advanced machine learning techniques, adaptive training strategies, and real-time processing are essential to improving fraud detection accuracy and protecting financial transactions from cyber threats.

Fraudulent transactions are rare, making them difficult to identify in large datasets. Traditional fraud-detection models often struggle because of class imbalance, where the number of legitimate transactions far outweighs fraudulent transactions. This imbalance leads to high false-negative rates, allowing fraud to remain undetected. To improve accuracy, adaptive training strategies**,** such as oversampling techniques such as Synthetic Minority Oversampling Technique (SMOTE), are

essential to ensure that machine learning models can effectively recognize fraudulent patterns. The primary goal of this study is to develop a fraud-detection model using Logistic Regression, a widely used machine-learning algorithm for binary classification. Because legitimate ones significantly outnumber fraudulent transactions, the SMOTE is applied to balance the dataset, ensuring that the model learns to recognize fraud effectively. Additionally, feature scaling and preprocessing enhance the model performance, leading to higher precision, recall, and ROC-AUC scores. This study focuses on binary classification, distinguishing between legitimate and fraudulent transactions. Machine learning techniques improve fraud detection accuracy while addressing dataset imbalances. The ultimate goal is to develop a scalable and efficient model capable of real-time fraud detection, making financial transactions more secure and reliable.

2. Literature Review

2.1 Overview of Fraud Detection

Fraud detection has traditionally relied on rule-based systems, in which predefined rules and thresholds are used to flag suspicious transactions[7]. These rules may include sudden high-value purchases, transactions from unusual locations, or multiple transactions within a short period[8]. Although effective in some cases, rule-based methods lack adaptability and often generate false positives, inconveniencing legitimate users. In contrast, machine-learning-based approaches analyze transaction data patterns and learn to detect fraud dynamically[9]. These models adapt to new fraud tactics and improve over time, thereby making them more accurate and efficient. Machine learning provides a scalable and intelligent fraud detection system by leveraging historical data and statistical techniques.

2.2 Machine Learning Models for Fraud Detection

2.2.1 Logistic Regression

Machine learning has become a powerful tool for detecting credit card fraud by identifying patterns in the transaction data. Unlike traditional rule-based methods, machine-learning models can adapt to changing fraud techniques and analyze large datasets efficiently. One commonly used model for fraud detection is logistic regression, which is a simple and interpretable algorithm for binary classification problems such as determining whether a transaction is fraudulent or legitimate.

Logistic regression estimates the probability that a given input belongs to a particular class by using the logistic (sigmoid) function. The model is represented by Eq. 1:

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}} \quad (1)$$

Where,

$P(Y=1|X)$ is the probability of the transaction being fraudulent.

β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_n$ are the weights associated with each feature $X_1, X_2, \ldots, X_n$

The function maps the input values to a range between 0 and 1, making it suitable for classification tasks.

One of the biggest challenges faced by logistic regression in fraud detection is the handling of unbalanced datasets. Because fraudulent transactions make up a tiny percentage of total transactions, the model favors the majority class (legitimate transactions), leading to high false-negative rates. This implies that actual fraud cases may go undetected. To address this issue, techniques such as the Synthetic Minority Oversampling Technique (SMOTE) can be used to balance the dataset by generating synthetic fraudulent samples. Additionally, adjusting the decision threshold of logistic regression can improve the sensitivity in detecting fraud. Logistic regression remains a widely used

model for fraud detection owing to its ease of implementation, computational efficiency, and ability to provide interpretable results.

2.3 Handling Imbalanced Datasets

One of the biggest challenges in fraud detection is dealing with imbalanced datasets, where fraudulent transactions comprise only a small percentage of the total transactions. If left unaddressed, this imbalance can cause machine-learning models to favor the majority class, leading to many false negatives. A common technique for addressing this issue is SMOTE, which generates synthetic data points for the minority class by interpolating existing samples. This helps create a more balanced dataset without simply duplicating existing cases. Another approach is undersampling, in which majority class samples are removed to balance the dataset. A hybrid method combines SMOTE and undersampling to maintain data integrity while improving the detection rates. In addition, feature engineering and preprocessing techniques such as scaling, outlier detection, and selecting meaningful transaction features can further enhance a model's ability to distinguish between fraudulent and legitimate transactions effectively[10].

2.3 Existing research works

Awoyemi et al. [10] highlighted the growing issue of financial fraud and the role of data mining in detecting credit card fraud. Their study compared Naïve Bayes, K-Nearest Neighbors (KNN), and Logistic Regression on an imbalanced dataset. Using hybrid sampling techniques, they evaluated model performance. Nuthalapati [11] emphasized the growing risk of credit card fraud and the need for machine learning-based detection frameworks. The study tackled imbalanced datasets using random undersampling and SMOTE to improve fraud detection. Comparing Random Forest and SVM, results showed Random Forest performed best, balancing precision and recall. The study highlighted the importance of continuous model adaptation against evolving fraud tactics. Lebichot et al. [12] highlighted the significant financial impact of credit card fraud despite its low occurrence. They emphasized the need for Fraud Detection Systems (FDS) capable of accurately identifying fraud while adapting to diverse fraud behaviors across payment systems and regions. The study stressed the importance of transfer learning to improve fraud detection across different financial contexts. Pozzolo et al. [13] explored credit card fraud detection as a complex challenge for computational intelligence, highlighting issues like concept drift, class imbalance, and verification latency. They emphasized the real-world limitations of existing fraud detection models and proposed a new learning strategy to tackle these challenges. Their experiments on 75 million transactions demonstrated the impact of class imbalance and evolving fraud patterns**.**

2.5 Research Gap

Logistic regression is widely used for fraud detection, but it struggles with imbalanced datasets, often misclassifying fraudulent transactions as legitimate transactions. Without proper preprocessing, performance is significantly affected. To improve accuracy, adaptive training strategies such as SMOTE are necessary to balance the dataset and enhance fraud detection.

3. Methodology

3.1 Dataset Description

The dataset used in this study was sourced from Kaggle's Credit Card Fraud Detection Dataset, which is widely used in machine-learning-based fraud detection research[14]. It contains anonymized credit card transactions, making it a reliable benchmark for evaluating fraud-detection models. The dataset comprises 568,630 transactions, each represented by 28 anonymized features obtained through Principal Component Analysis (PCA). These features capture the transaction behavior while ensuring data privacy. Additionally, the dataset includes amount, representing the transaction value, and class, which indicates whether a transaction is fraudulent or legitimate. A significant challenge in using this dataset is its highly imbalanced nature, as fraudulent transactions account for only 0.2% of total records. This class imbalance can lead to biased machine-learning

models that favor legitimate transactions unless proper resampling techniques are applied. Table 1 summarizes the key attributes of the dataset.

Table 1. Description of dataset

Attribute	**Details**
Source	Kaggle Credit Card Fraud Dataset
Total Transactions	568,630
Legitimate Transactions	567,493 (99.8%)
Fraudulent Transactions	1,137 (0.2%)
Features	28 PCA-transformed features
Additional Fields	Amount, Class (0: Legitimate, 1: Fraudulent)

Owing to the highly imbalanced distribution, techniques like SMOTE the Synthetic Minority Oversampling Technique (SMOTE) are necessary to enhance the model performance by generating synthetic fraudulent samples. Addressing this imbalance ensures better fraud detection and reduces false-negative rates, thereby making the model more effective in real-world applications.

3.2 Data Preprocessing

Appropriate data preprocessing is crucial for improving the accuracy and reliability of fraud-detection models. Before training the model, several pre-processing steps were applied to ensure data quality and balance.

3.2.1 Handling Missing Values

A thorough examination of the dataset revealed no missing values. This eliminated the need for imputation, allowing the focus to remain on feature scaling and class balancing to enhance the model performance.

3.2.2 Feature Scaling

Because the dataset contains both PCA-transformed features and a numerical transaction amount field, StandardScaler was applied to normalize the values. This transformation ensures that all features contribute equally to the model, preventing bias toward larger numerical values such as transaction amounts. Standardizing the data improves the efficiency of machine learning algorithms and helps logistic regression to perform optimally.

3.2.3 Class Balancing

The dataset was highly imbalanced, with fraudulent transactions accounting for only 0.2% of the total records. To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was used to generate synthetic fraudulent transactions and balance the dataset. This approach prevents the model from favoring legitimate transactions and enhances its ability to identify fraud correctly. By applying SMOTE, the model learns to detect fraudulent patterns more effectively, reducing false-negative rates and improving recall for fraudulent transactions. These preprocessing steps ensure that the dataset is well-structured, making the fraud detection model more reliable and accurate.

3.3 Model Design

The fraud detection model in this study is based on logistic regression, which is a widely used algorithm for binary classification. Logistic regression was chosen because of its simplicity, interpretability, and efficiency, making it well suited for distinguishing between legitimate and fraudulent transactions. Although logistic regression performs well, it requires proper preprocessing to handle class imbalance and ensure better fraud-detection accuracy.

3.3.1 Adaptive Training Strategies

Because fraudulent transactions are rare, the dataset is highly unbalanced. To address this, SMOTE is used to generate synthetic fraudulent samples, ensuring that the model learns fraud patterns more effectively. Additionally, undersampling of the majority class (legitimate transactions) was

considered an optional step to balance the dataset further, preventing the model from being biased toward the majority class.

3.3.2 Model Training & Evaluation

The dataset was split into 80% training and 20% testing datasets to ensure a well-distributed evaluation. The Saga solver was selected for logistic regression because it was optimized for large datasets and ensured faster convergence.

Several key metrics are used to assess the performance of the model.

- Accuracy, which measures the overall accuracy
- Precision, Recall, and F1-score evaluate the model's ability to correctly identify fraudulent transactions.
- The ROC-AUC score indicates how well the model distinguishes between fraud and non-fraud cases.

For better visualization, tools such as the confusion matrix, ROC curve, and precision-recall curve were used to analyze the effectiveness and performance of the model in detecting fraud.

4. Results and Discussion

4.1 Model Performance

4.1.1 Confusion Matrix Analysis

A confusion matrix provides a detailed evaluation of the performance of a classification model by comparing predicted values with actual labels. This helps identify how well the model distinguishes between fraudulent and legitimate transactions. The confusion matrix from the present study is shown in Fig. 1 and Table 2. In the fraud detection model, the confusion matrix is expressed as follows:

- **True positive (TP): 54,108** – Fraudulent transactions correctly identified as fraud.
- **True Negatives (TN): 55,625** – Legitimate transactions correctly classified as non-fraudulent.
- **False positive (FP): 1,238** – legitimate transactions incorrectly flagged as fraud.
- **False Negatives (FN): 2,755** – Fraudulent transactions incorrectly classified as legitimate.

A high number of true positives and negatives indicates that the model effectively detects fraud and non-fraud cases. However, false negatives (2,755 instances) remain a concern, as they represent fraudulent transactions that went undetected. Reducing false negatives is crucial for minimizing financial losses. Similarly, false positives (1,238 cases) indicate legitimate transactions mistakenly flagged as fraud, which can inconvenience customers. A well-balanced fraud-detection model aims to reduce false negatives and positives, while maintaining high precision and recall. Overall, the confusion matrix highlights the effectiveness of the model with a strong ability to correctly classify most transactions. Further optimization, such as fine-tuning thresholds or using advanced ensemble models, can improve detection rates and reduce misclassification errors.

Table 2 predicts Legitimacy and Fraud from this study.

	Predicted Legitimate (0)	**Predicted Fraudulent (1)**
Actual Legitimate (0)	55,625	1,238
Actual Fraudulent (1)	2,755	54,108

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 3, March : 2025

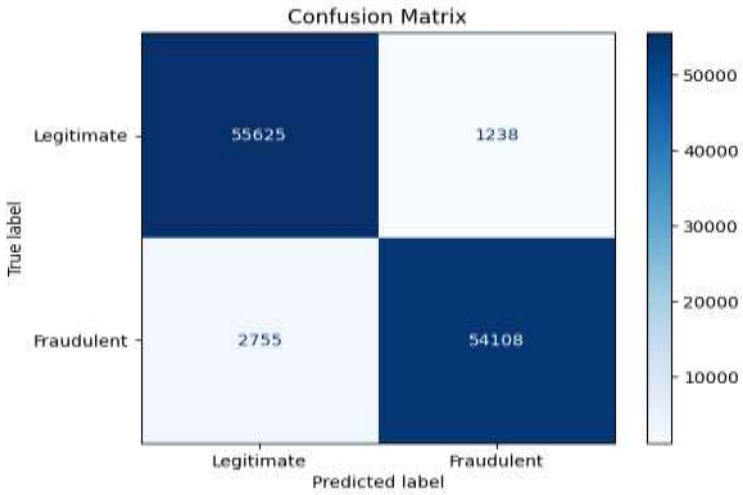

Fig. 1 Confusion matrix of credit card fraud prediction

4.1.2 Classification Report Analysis

The classification report comprehensively evaluates the performance of the fraud detection model. The model was assessed based on precision, recall, F1-score, accuracy, and ROC-AUC score to determine its effectiveness in identifying fraudulent transactions.

Table 3 Classification report from present study

Metric	Legitimate (0)	Fraudulent (1)	Macro Average	Weighted Average
Precision	0.95	0.98	0.97	0.97
Recall	0.98	0.95	0.96	0.96
F1-Score	0.97	0.96	0.96	0.96

I. Precision, Recall, and F1-Score

i. **Precision (Fraudulent Transactions): 98%**

Precision measures how many of the predicted fraudulent transactions are actually fraudulent. A precision score of 98% means that most flagged fraud cases were indeed fraudulent, minimizing false positives and reducing the inconvenience for legitimate users.

ii. **Recall (Fraudulent Transactions): 95%**

Recall measures the number of fraudulent transactions correctly identified. A recall score of 95% shows that the model successfully detected the most fraudulent activities while only missing a small percentage.

iii. **F1-Score: 96%**

The F1-score is the harmonic mean of precision and recall. A score of 96% reflects a well-balanced model that effectively identifies fraud while keeping the false positives low.

II. Overall Accuracy and ROC-AUC Score

i. **Accuracy Score: 96.49%**

The model correctly classified 96.49% of all transactions, demonstrating strong overall performance. However, because accuracy alone can be misleading for imbalanced datasets, other metrics, such as recall and ROC-AUC, provide deeper insights.

ii. **ROC-AUC Score: 0.9935**

The Receiver Operating Characteristic - Area Under Curve (ROC-AUC) score of 0.9935 indicates that the model is highly effective at distinguishing between fraudulent and legitimate transactions. A score close to 1.0 confirms that the model has strong discriminatory power.

The high precision and recall of the model indicate that it is highly reliable for fraud detection. The balance between accuracy, recall, and ROC-AUC ensures that fraudulent transactions are effectively identified while minimizing false positives.

4.2 Comparison with Other Models

Fraud detection relies on machine-learning models that accurately distinguish between legitimate and fraudulent transactions. To evaluate the effectiveness of logistic regression, its performance was compared with other traditional machine learning models, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and naïve Bayes, as shown in Table 4.

4.2.1 Performance Analysis

The comparison table highlights key performance metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC score for each model.

i. Logistic Regression achieved an accuracy of 96.49%, with a high precision of 98% and recall of 95%, demonstrating its reliability in detecting fraud while minimizing false positives.
ii. The SVM performed slightly lower, with an accuracy of 95.8%, offering good balance but requiring more computational resources.
iii. KNN had an accuracy of 94.5% but struggled with imbalanced data, resulting in a lower recall.
iv. Naïve Bayes had the lowest accuracy at 92.0%, as its assumption of feature independence limited its ability to capture complex fraud patterns.

While all models showed reasonable performance, logistic regression outperformed the others in terms of overall accuracy and fraud detection efficiency. A high ROC-AUC score of 0.9935 indicates its strong ability to separate fraudulent and legitimate transactions. Given its simplicity, interpretability, and effectiveness, logistic regression remains a strong choice for fraud detection, particularly when combined with techniques such as SMOTE and feature scaling to handle class imbalance.

Table 4 Results comparison of Logistic Regression with other techniques

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	**96.49%**	**98%**	**95%**	**96%**	**0.9935**
SVM	~95.8%	95%	94%	94.5%	~0.9900
KNN	~94.5%	92%	91%	91.5%	~0.9700
Naïve Bayes	~92.0%	90%	88%	89%	~0.9600

4.3 Discussion on results and practical implications

Handling class imbalance is one of the most significant challenges in fraud detection because fraudulent transactions make up only a small fraction of the total dataset. Without proper balancing techniques, machine-learning models tend to favor the majority class, leading to high false-negative rates. To address this issue, SMOTE was applied to generate synthetic samples for fraudulent transactions. Applying SMOTE significantly improves the model's ability to detect fraudulent transactions. By balancing the dataset, the recall for fraud cases increased, meaning that fewer fraudulent transactions were undetected. This improvement was essential in reducing false negatives and was critical in fraud detection to minimize financial losses. Additionally, feature scaling ensures that the transaction amount and other features contribute equally to the model's predictions. Logistic regression, when combined with SMOTE and feature scaling, achieved an accuracy of 96.49%, a precision of 98%, and a recall of 95%, striking a strong balance between detecting fraud and minimizing false alarms.

The combined effect of SMOTE and feature scaling ensured that the model did not disproportionately favor legitimate transactions while maintaining interpretability and efficiency.

These results highlight the importance of preprocessing techniques for enhancing the performance of machine-learning models for fraud detection. The balanced precision and recall of the model make it reliable for real-world fraud prevention systems.

5. Conclusion

This study demonstrates that logistic regression, when combined with appropriate preprocessing techniques, is an effective method for credit card fraud detection. Despite its simplicity, the logistic regression performed well in identifying fraudulent transactions, achieving high accuracy, precision, and recall. However, because fraud datasets are highly imbalanced, the Synthetic Minority Oversampling Technique (SMOTE) is crucial for improving a model's ability to detect fraudulent cases. By balancing the dataset, SMOTE significantly enhances recall, reducing the number of fraudulent transactions misclassified as legitimate. The model achieved high precision (98%) and recall (95%), ensuring a low false-positive rate, which is essential for fraud detection to minimize unnecessary transaction blocks. The ROC-AUC score of 0.9935 confirms the strong ability of the model to differentiate between fraudulent and legitimate transactions. While the results are promising, real-time fraud detection remains challenging because transactions must be flagged immediately. Future work should focus on integrating real-time detection systems capable of instantly processing transactions. Additionally, exploring advanced deep learning models, such as neural networks, or ensemble methods, such as XGBoost, can enhance accuracy and adaptability. Implementing such improvements will help create a more robust, scalable, and efficient fraud detection system for real-world applications.

References

[1] B. Ul, R. F., A. Mehraj, A. Ahmad, and S. Assad, "A Compendious Study of Online Payment Systems: Past Developments, Present Impact, and Future Considerations," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 5, pp. 256–271, 2017, doi: 10.14569/ijacsa.2017.080532.

[2] Z. Bezhovski, "The Future of the Mobile Payment as Electronic Payment System," *Eur. J. Bus. Manag.*, vol. 8, no. 8, pp. 127–132, 2016.

[3] J. Kang, "Mobile payment in Fintech environment: trends, security challenges, and services," *Human-centric Comput. Inf. Sci.*, vol. 8, no. 1, 2018, doi: 10.1186/s13673-018-0155-4.

[4] Sumanjeet, "Emergence of payment systems in the age of electronic commerce: The state of art," in *2009 First Asian Himalayas International Conference on Internet*, 2009, pp. 1–18. doi: 10.1109/AHICI.2009.5340318.

[5] B. K. Nambiar and K. Bolar, "Factors influencing customer preference of cardless technology over the card for cash withdrawals: an extended technology acceptance model," *J. Financ. Serv. Mark.*, vol. 28, no. 1, pp. 58–73, 2023, doi: 10.1057/s41264-022-00139-y.

[6] V. Komandla, "The Digital Wallet Revolution : Adoption Trends , Consumer Preferences , and The Digital Wallet Revolution : Adoption Trends , Consumer Preferences , and Market Impacts on Bank-Customer Relationships," no. April 2018, 2024.

[7] D. Birch and M. A. Young, "Financial services and the Internet - what does cyberspace mean for the financial services industry?," *Internet Res.*, vol. 7, no. 2, pp. 120–128, Jan. 1997, doi: 10.1108/10662249710165262.

[8] B. Świecka, P. Terefenko, and D. Paprotny, "Transaction factors' influence on the choice of payment by Polish consumers," *J. Retail. Consum. Serv.*, vol. 58, p. 102264, 2021, doi: https://doi.org/10.1016/j.jretconser.2020.102264.

[9] D. L. Shrier, *Digital Financial Services*, no. April. 2022. doi: 10.7551/mitpress/13673.003.0008.

[10] J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare, "Credit card fraud detection using machine learning techniques: A comparative analysis," in *2017 International Conference on*

Computing Networking and Informatics (ICCNI), 2017, pp. 1–9. doi: 10.1109/ICCNI.2017.8123782.
[11] A. Nuthalapati, "Smart Fraud Detection Leveraging Machine Learning For Credit Card Security," *Educ. Adm. Theory Pract.*, vol. 29, no. 2, pp. 433–443, 2023, doi: 10.53555/kuey.v29i2.6907.
[12] B. Lebichot, Y.-A. Le Borgne, L. He-Guelton, F. Oblé, and G. Bontempi, "Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection BT - Recent Advances in Big Data and Deep Learning," L. Oneto, N. Navarin, A. Sperduti, and D. Anguita, Eds., Cham: Springer International Publishing, 2020, pp. 78–88.
[13] A. D. Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi, "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 8, pp. 3784–3797, 2018, doi: 10.1109/TNNLS.2017.2736643.
[14] A. D. Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," *Proc. - 2015 IEEE Symp. Ser. Comput. Intell. SSCI 2015*, no. November, pp. 159–166, 2015, doi: 10.1109/SSCI.2015.33.