

Journal of Hunan University (Natural Sciences)

Vol. 53 No. 4

April 2026

Available online at

<https://jonuns.com>



ELSEVIER
Scopus



Clarivate
WEB OF SCIENCE

Open Access Article

 <https://doi.org/10.55463/issn.1674-2974.53.4.1>

Improving Fraud Detection Systems Using Deep Learning with Resampling Techniques

Abdul Ahad Hassan Farroqi ^{1*}, Shah Ameer ², DR. Zahoor Ahmad ², Fahad Hassan Farooqi ³

¹ School of Economics and Trade, Hunan university, Changsha, China,

² Department of Statistics, University of Sargodha, Sargodha, Pakistan,

³ Department of Statistics, Government Graduate College, Jhang, Pakistan,

* Corresponding author: abdulahadfaroqi73@gmail.com

Article History:

Received: February 5, 2026

Revised: March 20, 2026

Accepted: April 6, 2026

Published: April 25, 2026

Abstract: Credit card fraud detection remains challenging because real transaction data are extremely imbalanced, where fraudulent cases represent only a tiny fraction of total observations. Models trained on such skewed data can achieve very high overall accuracy while still failing to detect fraud reliably, limiting practical usefulness. This study investigates whether resampling-assisted deep learning can improve minority-class (fraud) detection without generating an impractical false-alarm burden. Using the publicly available Kaggle Credit Card Fraud Detection dataset (284,807 transactions with 492 fraud cases), we evaluate three deep learning architectures—Multilayer Perceptron (MLP), Deep Belief Network (DBN), and Convolutional Neural Network (CNN)—under three training settings: no resampling, Random Under-Sampling (RUS), and Synthetic Minority Over-Sampling Technique (SMOTE). The novelty of this work lies in a controlled comparison of SMOTE versus RUS across multiple deep architectures under a consistent preprocessing pipeline, where resampling is applied only to the training set to prevent information leakage and preserve realistic testing. Model performance is assessed using accuracy together with precision, recall, and F1-score to reflect rare-event detection priorities. The results show that RUS increases fraud recall (0.90–0.92) across models but yields very low precision (0.03) and low F1-scores (0.05–0.07), indicating



Copyright: © 2026 by the authors. Licensee JHU

This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>)

excessive false positives that reduce deployment feasibility. In contrast, SMOTE produces a balanced improvement in fraud detection while maintaining very high accuracy. DBN + SMOTE achieves the best overall balance with 99.89% accuracy, 0.63 precision, 0.86 recall, and the highest F1-score of 0.73, while MLP + SMOTE achieves the highest accuracy (99.90%) with 0.59 precision, 0.86 recall, and F1-score of 0.70; CNN + SMOTE also performs competitively (99.85% accuracy, 0.54 precision, 0.81 recall, F1-score 0.65). These findings demonstrate that SMOTE-assisted deep learning provides a more deployable precision-recall trade-off than RUS for imbalanced fraud detection, improving fraud recognition while controlling false alarms.

Keywords: Fraud detection, Deep learning, SMOTE, Random under-sampling (RUS), Imbalanced classification.

基于重采样技术的深度学习信用卡欺诈检测方法研究

摘要：信用卡欺诈检测面临的核心困难在于真实交易数据高度不平衡：欺诈样本极少，但漏检可能造成严重经济损失与风险。由于类别分布偏斜，模型即使获得很高的总体准确率，也可能无法有效识别欺诈交易，从而影响实际应用价值。本研究旨在验证“重采样 + 深度学习”是否能够在提升少数类（欺诈）识别能力的同时避免产生不可接受的误报负担。基于公开的 Kaggle 信用卡欺诈检测数据集（共 284,807 笔交易，其中 492 笔为欺诈），本文在统一的预处理流程下比较三种深度学习模型：多层感知机（MLP）、深度置信网络（DBN）与卷积神经网络（CNN），并在三种训练设置下进行实验：不重采样、随机欠采样（RUS）与合成少数类过采样技术（SMOTE）。本文的创新点在于：在相同数据集与相同预处理条件下，对 SMOTE 与 RUS 在多种深度架构上的影响进行受控对比，并且仅在训练集上进行重采样以避免信息泄漏，从而保证测试更贴近真实部署场景。评价指标采用准确率、精确率、召回率与 F1-score，以反映稀有事件检测的实际需求。实验结果表明，RUS 可显著提高召回率（0.90-0.92），但精确率极低（0.03）且 F1-score 较低（0.05-0.07），说明误报大量增加，不利于实际监控系统部署。相较之下，SMOTE 在保持极高准确率的同时实现更均衡的欺诈检测性能。其中 DBN + SMOTE 表现最佳：准确率 99.89%，精确率 0.63，召回率 0.86，F1-score 达到最高的 0.73；MLP + SMOTE 准确率最高（99.90%），同时保持 0.59 的精确率、0.86 的召回率与 0.70 的 F1-score；CNN + SMOTE 也具有竞争力（准确率 99.85%，精确率 0.54，召回率 0.81，F1-score 0.65）。综上，本文证明了在极端不平衡的信用卡欺诈检测任务中，SMOTE 辅助的深度学习方法比 RUS 更能在“提高欺诈识别”与“控制误报成本”之间取得可部署的平衡。

关键词： 欺诈检测；深度学习；SMOTE；随机欠采样（RUS）；类别不平衡分类

1. Introduction

Modern financial activity has rapidly expanded with electronic commerce and digital banking, making credit card payments a core component of consumer and business transactions. Alongside these benefits, credit card fraud has become a persistent and costly problem for financial institutions and customers. Fraudulent behavior is adaptive and concealed within very large volumes of legitimate transactions; therefore, manual inspection is not practical and automated fraud detection

systems are essential. Recent studies show that machine learning approaches can discover hidden transactional patterns and support time-efficient fraud screening, offering advantages over rigid rule-based systems that are difficult to maintain and may fail as fraud strategies evolve [1,4,5]. In real operational settings, fraud detection models must balance two competing goals: detecting as many fraudulent events as possible (high recall) while keeping false alarms manageable to avoid overwhelming investigators and customer service workflows [2,10].

A central difficulty in credit card fraud detection is the extreme class imbalance present in real transaction data, where fraud cases form only a small fraction of all observations. When classifiers are trained on such imbalanced data, they tend to favor the majority (legitimate) class. As a result, a model may report very high overall accuracy while still missing many fraud cases, which is risky in financial applications. This concern is well established in the imbalanced learning literature, which emphasizes that training strategies and evaluation measures must be designed carefully so that minority-class performance is not underestimated or hidden by overall accuracy [8,9]. For this reason, fraud detection performance should be reported using minority-focused measures such as precision, recall, and F1-score, where precision reflects false-alarm cost and recall reflects missed-fraud risk [14,17]. Moreover, PR-based evaluation is often more informative than ROC-based interpretation under heavy imbalance, because it better reflects the positive-class retrieval task that investigators care about [14,17].

To address imbalance, many studies apply data-level resampling methods. SMOTE is one of the most widely used approaches; instead of duplicating minority samples, it generates synthetic minority instances through interpolation to enhance minority representation and improve rare-event detection [7]. Another common strategy is random under-sampling (RUS), which reduces majority dominance by removing a subset of legitimate transactions. However, under-sampling may discard informative majority examples and can increase false alarms depending on the classifier and the structure of the feature space [8,9,15]. Consequently, the choice between SMOTE and RUS is not purely methodological; it can change the operational trade-off between fraud recall improvement and the cost of investigating additional false positives [2,15].

Deep learning has gained popularity in fraud detection because it can model nonlinear relationships and complex feature interactions that may be difficult for traditional methods to capture. CNN-based architectures and other neural approaches have been explored for transaction classification, indicating that representation learning can be applied beyond images to structured numerical features [12]. Recent peer-reviewed studies further suggest that deep learning models can achieve strong fraud detection results on benchmark datasets, particularly when combined with imbalance mitigation strategies [2,3,13]. At the same time, surveys of the field highlight that reported performance varies widely across papers, partly because experimental protocols differ (feature scaling, data splitting, resampling configuration, threshold selection, and metric reporting) [5,10]. This variability motivates controlled evaluations where multiple models are compared under a single consistent pipeline, enabling

more reliable conclusions about which design choices matter most [5,10].

Motivated by these challenges, this paper evaluates a resampling-assisted deep learning framework that combines SMOTE and RUS with three neural models: Multilayer Perceptron (MLP), Deep Belief Network (DBN), and Convolutional Neural Network (CNN). Model performance is assessed using accuracy along with precision, recall, and F1-score to provide a comprehensive view of fraud detection performance [8,9,14]. The primary research question is: under the same dataset and preprocessing pipeline, which resampling strategy (SMOTE or RUS) yields more reliable minority-class detection across different deep architectures? By answering this question, the study clarifies whether minority enrichment (SMOTE) or majority reduction (RUS) is more effective for deep learning on transaction data, and whether this effect is consistent across architectures.

The novelty of this work lies in providing a controlled, architecture-level comparison of resampling strategies rather than evaluating a single model in isolation. Specifically, we (i) evaluate three deep architectures under identical preprocessing and train/test splitting, (ii) apply resampling only to the training set to avoid leakage and preserve realistic evaluation conditions, and (iii) emphasize minority-focused metrics to reflect deployment priorities. These design choices address common reproducibility and comparability limitations reported in prior studies and reviews [5,10].

The main contributions of this study are:

- (1) A comparative evaluation of MLP, DBN, and CNN for credit card fraud detection under severe class imbalance.
- (2) An empirical analysis of SMOTE versus RUS on minority-class performance grounded in imbalanced learning principles [7–9,15];
- (3) Discussion of results in the context of recent fraud detection research and current trends in deep learning and imbalance mitigation [1–6,10–13].
- (4) Deployment-oriented guidance by interpreting metric changes (precision/recall/F1) in terms of investigation workload and missed-fraud risk.

The remainder of this paper is organized as follows. Section 2 describes the dataset, preprocessing, resampling strategies, models, and evaluation metrics. Section 3 presents results and discussion. Section 4 concludes the paper and suggests directions for future work.

1.1 Literature Review

1.1.1 Fraud detection approaches: rules and machine learning

Early and practical fraud detection systems often relied on expert rules and domain heuristics, which can be interpretable and easy to deploy but are difficult to maintain as fraud strategies evolve. Semantic and rule-based approaches remain useful for expressing business logic and compliance constraints, yet their rigidity can limit adaptability when fraud behavior changes rapidly [4]. Machine learning methods have therefore become widely used for fraud detection because they learn patterns from historical transactions and can scale to large volumes of data [1,5,6]. In addition to accuracy, many modern fraud detection studies emphasize operational constraints such as false-alarm rates and the cost of investigation, because excessive alerts can degrade real-world usefulness even when recall is high [2,10]. Chang et al. [1] provide evidence on the Kaggle benchmark dataset that model choice and imbalance handling can materially affect fraud-detection outcomes, supporting the need for careful experimental design and evaluation.

1.1.2 Deep learning for credit card fraud detection

Deep learning models can capture nonlinear feature interactions and may learn richer representations than traditional classifiers. Neural networks have been explored for transaction classification, and CNN-based methods demonstrate that convolutional representation learning can be applied to structured numerical features by organizing inputs into learnable local patterns [12]. Peer-reviewed studies have reported strong performance using deep learning models for fraud detection when combined with appropriate preprocessing and training strategies, including imbalance mitigation [2,3,13]. Wu et al. [3], for example, investigate a deep learning approach on credit card fraud detection and report that network design choices can influence the precision–recall trade-off, which is critical in rare-event screening.

However, surveys consistently observe that results across papers are difficult to compare because experimental protocols differ particularly the handling of class imbalance, the use of resampling on the full dataset (which can cause leakage), and differences in threshold selection and metric reporting [5,10]. As a result, controlled evaluations that standardize these choices are needed to distinguish true methodological improvements from protocol-dependent gains.

1.1.3 Handling class imbalance: resampling and its trade-offs

Imbalanced learning research emphasizes that naïve training on skewed distributions biases models toward the majority class, motivating resampling, cost-sensitive learning, and other mitigation strategies [8,9]. Among resampling techniques, SMOTE is widely adopted because it increases minority representation by generating synthetic samples through interpolation, helping reduce majority bias while preserving minority diversity [7]. Under-sampling methods such as RUS balance classes by removing a subset of majority samples, which can reduce training time and mitigate dominance effects but may also remove informative examples that help define decision boundaries [8,9,15]. Therefore, SMOTE and RUS can lead to different operational outcomes: SMOTE may increase recall by enriching fraud examples, while RUS may change the classifier boundary in a way that either helps detect fraud or increases false positives depending on the retained majority structure [8,9,15]. Because these effects can interact with model capacity and representation learning, it is important to evaluate resampling strategies across multiple deep architectures rather than assuming consistent benefits across models.

1.1.4 Evaluation of rare-event classifiers: PR vs ROC and metric selection

Evaluation methodology is critical in fraud detection because the positive class is rare and the cost of errors is asymmetric. Davis and Goadrich [14] formalize the relationship between precision–recall and ROC curves, highlighting how the same classifier behavior can be interpreted differently under class imbalance. Saito and Rehmsmeier [17] further argue that precision–recall analysis is often more informative than ROC analysis when evaluating imbalanced binary classification. In fraud detection, precision relates to the burden of investigating false alarms, while recall relates to missed fraud risk, so both metrics must be reported together rather than using accuracy alone [14,17]. Accordingly, many imbalanced-learning surveys recommend emphasizing PR-based metrics and F1-score for rare-event tasks, especially when comparing imbalance mitigation techniques [8,9,17].

Although prior work supports deep learning and resampling for credit card fraud detection, fewer studies provide a controlled comparison across multiple deep architectures while directly contrasting SMOTE versus RUS under identical preprocessing and evaluation settings, with resampling confined to the training data to avoid leakage. This gap motivates the comparative framework examined in this study and supports the novelty and practical relevance highlighted by the reviewers.

2. Methods and Materials

2.1 Dataset Description

In this work, I will be employing the publicly available Kaggle Credit Card Fraud Detection data that consists of the 284,807 credit card transactions conducted by European customers in September 2013. The number of transactions classified as fraudulent is only 492, demonstrating a very strong class imbalance that characterizes real-world fraud detection tasks. The data were formatted to be confidential: the majority of attributes have been processed through Principal Component Analysis (PCA), and presented in anonymized numerical components V1–V28. Two variables are unaltered, namely Time and Amount. The Amount variable is the transaction value and Time is the time taken (in seconds) separating any given transaction and the initial transaction in a series of transactions. The dependent variable Class is dichotomous with Class = 1 showing fraud and Class = 0 showing a legitimate transaction.

This dataset was selected because it is a widely used benchmark for credit card fraud research, exhibits severe real-world class imbalance, and contains anonymized PCA features that reflect practical privacy constraints in financial data. The same dataset has also been used in prior work (e.g., Chang et al. [1]), which supports meaningful comparison of findings across studies.

Since the data contain anonymized PCA components, the focus of the model structure is the acquisition of reliable decision boundaries based on numerical regularities as opposed to interpretation of original transaction attributes. Also, the severe imbalance implies that assessment should be concerned with minority-class detection as opposed to general accuracy.

Practical implication is that in deployment, models must detect rare fraud cases while controlling false alarms; therefore, improvements must be interpreted using minority-focused metrics rather than accuracy alone.

2.2 Data Preprocessing

A stable preprocessing pipeline was used to compare and make models stable during training. First, all input features were normalized with Z-score normalization, so that all features have similar scale. Neural networks in particular benefit from standardization because it improves numerical conditioning and may aid convergence in optimization.

The dataset was split into an 80/20 training and testing set. In order to have realistic assessment and prevent information leakage, resampling to change the distribution of classes (i.e., SMOTE and RUS) was applied only to the training portion, whereas the test

portion was left in its initial imbalanced form to be evaluated in terms of final performance.

The dataset was normalized and then used for training under imbalanced and resampled conditions. An important methodological constraint of imbalanced learning is that it should not be leaked during learning; hence, resampling methodology was only used on the training sample, and the test sample was not resampled. This maintains realism of assessment since, in a real-world implementation, fraud is rare and the model must perform under an inherently skewed distribution.

To improve reproducibility, the same split ratio, normalization method, and resampling-only-on-training rule were applied consistently for all three deep architectures.

Two imbalance-handling techniques were analyzed: Synthetic Minority Over-Sampling Technique (SMOTE) and Random Under-Sampling (RUS). The choice of these methods is inspired by their different mechanisms and popularity: oversampling enhances minority representation without discarding majority samples, while under-sampling reduces majority dominance by removing a portion of the majority class. A comparison of the two strategies can give insight into whether minority augmentation or majority reduction plays a stronger role in learning behavior of deep architectures on transaction data.

2.3 Resampling Techniques

2.3.1 Synthetic Minority Over-Sampling Technique (SMOTE)

SMOTE is a type of oversampling meant to enhance learning of rare classes by generating synthetic minority examples. Rather than copying existing fraud samples, SMOTE creates new minority examples by interpolating between a minority sample and a selected set of nearest-neighbor minority samples in feature space. This enlarges minority samples and preserves feature diversity, which can reduce overfitting compared to simply replicating the minority class. This is especially significant in fraud detection because the model must learn fraud patterns based on a limited number of original examples. The practical purpose of applying SMOTE in the present research is to provide models with a larger range of fraud-like samples during training to enhance minority-class recognition.

In this study, SMOTE is evaluated to determine whether minority enrichment improves recall without causing an unacceptable decrease in precision (i.e., excessive false alarms).

2.3.2 Random Under-Sampling (RUS)

RUS minimizes imbalance by randomly sampling a subset of majority-class (legitimate) transactions so that the resultant training sample is more balanced. Under-sampling can decrease training time and reduce majority bias, particularly when the majority

class is very large compared to the minority class. Nevertheless, since it removes data, RUS can eliminate informative legitimate patterns that are needed to train strong boundaries. Therefore, its efficacy cannot be guaranteed and must be evaluated empirically. RUS is used in this study both to provide a direct comparison with SMOTE and to determine whether reducing majority dominance leads to improved fraud recall or reduced generalization.

In this study, RUS is evaluated to quantify the trade-off between potential recall gains and potential precision loss due to majority information removal.

2.4 Deep Learning Models

Three deep learning networks were tested: Multilayer Perceptron (MLP), Deep Belief Network (DBN), and Convolutional Neural Network (CNN). The choice of these models was motivated by the fact that they embody different learning principles and help determine whether the same resampling strategy benefits different architectures in similar or different ways.

This multi-architecture design supports the novelty of the study by providing a controlled comparison rather than reporting results from only a single model.

2.4.1 Multilayer Perceptron (MLP)

MLP is a supervised feedforward neural network that learns nonlinear mappings from input features to the target label. It is a robust foundation for tabular data and is widely employed in fraud detection since it can capture complex feature interactions without manual feature engineering. The MLP provides a benchmark to determine the effect of resampling on a typical dense-network classifier.

2.4.2 Deep Belief Network (DBN)

A DBN is a deep architecture conventionally built from stacked Restricted Boltzmann Machines and typically trained in two phases: unsupervised layer-wise pretraining followed by supervised fine-tuning. The reason to incorporate a DBN is that it can acquire hierarchical feature representations that may help differentiate subtle fraud patterns embedded in noisy transactional data. Although DBNs are less common in modern feedforward architectures, they remain of academic interest for studying representation learning under extreme imbalance.

2.4.3 Convolutional Neural Network (CNN)

CNNs are trained using convolutional filters and are effective in many domains. Though CNNs are typically applied to image and signal processing, they can be adapted to structured numerical data whereby local combinations of features are learned and refined into more abstract forms. CNN-based methods can

detect discriminative patterns not easily expressed as linear combinations of features in fraud detection. Including a CNN enables comparison between convolution-based representation learning and fully connected methods.

Experiments were carried out under three training setups for each architecture: (i) no resampling (original imbalanced training set), (ii) SMOTE applied to the training set, and (iii) RUS applied to the training set. This experimental design holds the preprocessing pipeline constant and ensures that performance differences can be attributed mainly to the resampling strategy and architecture rather than different preprocessing pipelines.

2.5 Evaluation Metrics

Since fraud detection is a rare-event classification problem, performance should be measured using metrics that reflect minority-class detection. Accuracy is not enough because a model may be accurate simply by predicting most transactions as legitimate due to the majority class. Consequently, accuracy, precision, recall, and F1-score are reported in this study to provide a balanced view.

Precision measures the proportion of predicted fraud cases that are truly fraud and reflects the false-alarm cost to investigators. Recall (sensitivity) measures the proportion of real fraud cases that are successfully detected; missed fraud (false negatives) may lead to direct monetary loss and reputational harm. F1-score combines precision and recall into a single measure and represents the trade-off between false positives and the ability to capture fraud. The joint interpretation of these metrics supports deployment-oriented evaluation and enables comparison of model behavior under different resampling strategies.

3. Results and Discussion

Detecting fraudulent transactions is challenging primarily because real credit card datasets are dominated by legitimate activity, whereas fraud constitutes a very small minority. Under this condition, overall accuracy can be misleading because a classifier can achieve a high accuracy by predominantly predicting the majority class while still failing to identify fraud cases. For this reason, the present study evaluates performance using accuracy together with minority-sensitive metrics (precision, recall, and F1-score), which more directly reflect fraud detection effectiveness. Three deep learning architectures MLP, DBN, and CNN were trained under three conditions: training on the original imbalanced data, training with Random Under-Sampling (RUS), and training with Synthetic Minority Over-Sampling Technique (SMOTE). To maintain realism and methodological validity, resampling was applied only to the training set, while the test set remained in its

naturally imbalanced distribution; this prevents information leakage and reflects real deployment where fraud remains rare.

3.1 Performance Comparison of Resampling Techniques

Across all architectures, the baseline models trained without resampling achieved very high accuracy but showed weak performance in detecting fraud. This outcome is consistent with imbalanced learning theory: when the fraud class is rare, the loss function is dominated by the majority class and the model is encouraged to optimize majority predictions, which inflates accuracy while reducing sensitivity to fraud. Consequently, the baseline results support the need to incorporate imbalance-handling strategies to strengthen minority-class learning.

When RUS was applied, fraud recall increased substantially for all models, reaching 92% for DBN (RUS), 92% for MLP (RUS), and 90% for CNN (RUS). However, these recall improvements occurred alongside very low fraud precision (0.03) and very low F1-scores (0.05–0.07). The combination of high recall and extremely low precision indicates that the RUS-trained models classified a large number of legitimate transactions as fraudulent. This behavior can be explained by the removal of majority-class examples during training, which reduces the model's exposure to the full variability of legitimate transactions and can lead to less specific decision boundaries. From an applied perspective, such a precision level would generate an excessive volume of false alerts, increasing the burden on investigators and potentially reducing the operational usefulness of the system. Therefore, while RUS improves sensitivity to fraud, it does so at a cost that is not consistent with practical fraud detection requirements where false alarms carry substantial operational and customer-service costs.

In contrast, SMOTE produced consistently strong and balanced performance across the three deep learning architectures. Under SMOTE, CNN achieved accuracy of 99.85%, fraud precision of 0.54, recall of 0.81, and F1-score of 0.65. DBN achieved accuracy of 99.89%, precision of 0.63, recall of 0.86, and the highest F1-score of the study (0.73). MLP achieved the highest accuracy (99.90%) with precision of 0.59, recall of 0.86, and F1-score of 0.70. These results suggest that SMOTE strengthens minority-class learning while preserving the information contained in the majority class, because it augments fraud examples without discarding legitimate observations. Accordingly, SMOTE yields a substantially better precision–recall balance than RUS and produces higher F1-scores, indicating improved overall effectiveness in rare-event detection. Taken together, Table 1 demonstrates that SMOTE-based training provides a more deployable performance profile than RUS-based training under the tested conditions.

Table 1: Model Performance Comparison (Compiled by the authors)

Model (Resampling)	Accuracy	Precision (Fraud)	Recall (Fraud)	F1-Score (Fraud)
DBN (RUS)	94.36%	0.03	0.92	0.05
MLP (RUS)	94.89%	0.03	0.92	0.06
CNN (RUS)	95.72%	0.03	0.90	0.07
CNN (SMOTE)	99.85%	0.54	0.81	0.65
DBN (SMOTE)	99.89%	0.63	0.86	0.73
MLP (SMOTE)	99.90%	0.59	0.86	0.70

As an overall finding, DBN + SMOTE provides the best balance between fraud detection and false-alarm control (highest F1-score), whereas MLP + SMOTE offers the highest accuracy with similarly strong recall. This difference reflects a meaningful operational choice: maximizing balanced performance (DBN + SMOTE) versus maximizing overall accuracy with strong but slightly less balanced fraud detection (MLP + SMOTE).

3.2 Confusion Matrix Analysis of the Best SMOTE-Based Models

To further interpret the practical behavior of SMOTE-based models beyond summary metrics, confusion matrices are examined for MLP + SMOTE and DBN + SMOTE. Confusion matrices explicitly show how many legitimate transactions are incorrectly flagged (false positives) and how many fraud cases are missed (false negatives), which is central to understanding deployment implications.

Figure 1 (MLP + SMOTE) shows that the model correctly identified 56,858 legitimate transactions and incorrectly flagged 6 legitimate transactions as fraud. For the minority class, the model correctly detected 74 fraud cases and missed 24 fraud cases. This indicates that the MLP + SMOTE configuration controls false positives very effectively while achieving strong fraud recall, which is desirable in scenarios where investigation capacity is limited and excessive alerts must be avoided. At the same time, the remaining false negatives indicate that some fraud cases still evade detection, suggesting room for further improvement through threshold calibration or cost-sensitive adjustments.

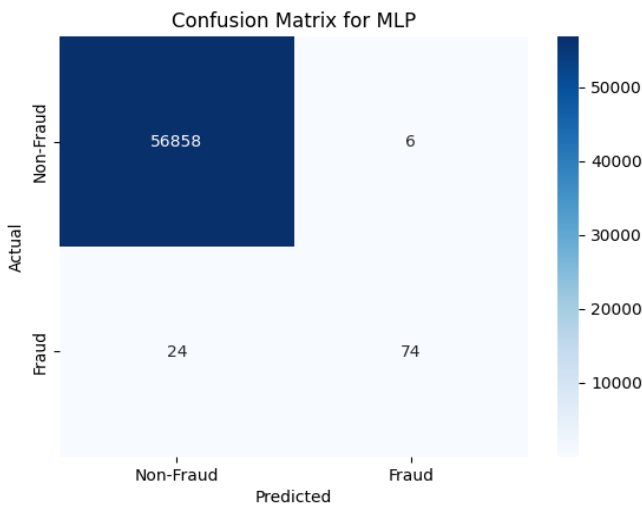


Fig. 1: Confusion matrix for MLP + SMOTE
(Developed by the authors)

Figure 2 (DBN + SMOTE) shows that the model correctly identified 56,815 legitimate transactions and incorrectly flagged 49 legitimate transactions as fraud. For fraud cases, the DBN correctly detected 84 and missed 14. Compared with MLP + SMOTE, DBN + SMOTE reduces missed fraud (false negatives) but increases false alerts (false positives), which explains its higher F1-score and reflects a different operating point on the precision–recall trade-off. This trade-off is important in applied fraud detection, where institutions may set different priorities depending on risk tolerance and available investigation resources.

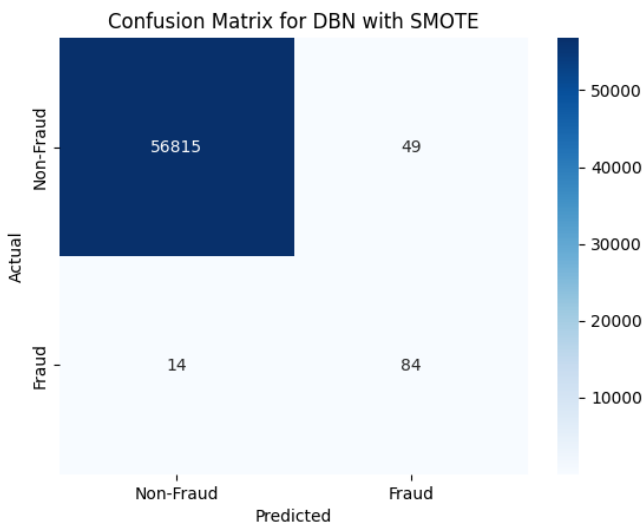


Fig. 2: Confusion matrix for DBN + SMOTE
(Developed by the authors)

3.3 Discussion in Relation to Prior Work and Implications

The results confirm that imbalance handling is a decisive factor in deep learning-based fraud detection. RUS increases recall by exposing the model to a more

balanced class distribution, but it also reduces the diversity of legitimate examples, which can destabilize the decision boundary and generate a large number of false positives. In practical systems, this behavior can be undesirable because it increases investigation workload and may reduce usability. In contrast, SMOTE maintains the full majority distribution while increasing minority visibility during training, leading to improved recall with substantially stronger precision and F1-score.

With respect to the broader literature, the present findings align with studies that emphasize the importance of imbalance mitigation for fraud detection performance. Chang et al. [1], using the same benchmark dataset, similarly indicate that model performance can change substantially when imbalance is appropriately addressed. Because differences in preprocessing, sampling configuration, and evaluation procedures can influence reported outcomes, the comparison is best interpreted qualitatively rather than as a direct numerical ranking. Nevertheless, both the prior work and the present results support the conclusion that imbalance-aware training improves fraud recognition in ways that accuracy alone does not capture. The present study contributes by providing a controlled comparison across three deep architectures under a single pipeline and by demonstrating that SMOTE yields a more favorable precision–recall balance than RUS in this setting.

Several limitations should be acknowledged. First, synthetic oversampling may increase overfitting risk if synthetic samples do not adequately reflect the true diversity of fraud patterns. Second, even the best-performing configuration still produced false negatives, implying that complete fraud capture is not achieved under the current setup. Future work could incorporate decision-threshold calibration, cost-sensitive learning, focal-loss training, or ensemble-based modeling to further reduce false negatives while maintaining acceptable false-positive rates. In addition, evaluating robustness across multiple random splits and considering calibration of predicted probabilities would strengthen evidence for deployment readiness.

Overall, the findings demonstrate that combining deep learning with resampling is an effective strategy for credit card fraud detection under extreme imbalance, and that SMOTE, in particular, offers a robust balance between fraud detection capability and false-alarm control on the benchmark dataset used in this study.

4. Conclusion

In this work, a deep learning approach for credit card fraud detection under severe class imbalance was investigated by comparing three architectures (MLP, DBN, and CNN) combined with two resampling strategies (RUS and SMOTE). The results confirm that class imbalance strongly affects model behavior: when

trained on the original imbalanced data, models can achieve very high overall accuracy while failing to reliably detect fraudulent transactions, demonstrating that accuracy alone is not sufficient for evaluating fraud detection systems and must be complemented by minority-focused metrics such as precision, recall, and F1-score.

Comparative experiments further demonstrate that the choice of resampling strategy is critical. Random Under-Sampling (RUS) produced very high fraud recall (90–92%) across the evaluated architectures, but it also resulted in extremely low precision (0.03) and very low F1-scores (0.05–0.07), indicating a large volume of false alarms that would be impractical for real-world monitoring and investigation workflows. In contrast, SMOTE consistently delivered a more balanced and deployable performance profile by improving fraud recall while maintaining very high accuracy and substantially better precision. Among SMOTE-based configurations, DBN + SMOTE achieved the best overall balance with the highest F1-score (0.73), while MLP + SMOTE achieved the highest accuracy (99.90%) with strong recall (0.86) and F1-score (0.70); CNN + SMOTE also produced competitive performance (F1 = 0.65). Overall, these results indicate that oversampling (SMOTE) is more appropriate than under-sampling (RUS) for resampling-assisted deep learning fraud detection on the benchmark dataset used in this study, because it improves minority detection without producing an excessive false-positive burden.

When considered in the context of prior work, the findings are consistent with studies that emphasize the importance of imbalance handling for fraud detection performance. In particular, Chang et al. [1], using the same Kaggle benchmark dataset, similarly highlight that appropriate imbalance treatment and minority-focused evaluation are necessary to obtain meaningful fraud detection performance. Although differences in preprocessing and evaluation protocols across studies limit direct numerical comparison, both the literature and the present results support the conclusion that imbalance-aware training improves fraud recognition beyond what overall accuracy alone can show. The present study extends this evidence by providing a controlled comparison of SMOTE versus RUS across three deep learning architectures under a single consistent pipeline, clarifying the practical precision–recall trade-offs.

From a practical perspective, the results imply that maximizing recall alone is not sufficient for deployment, because extremely low precision generates excessive false alarms and increases operational cost. Therefore, deployable fraud detection models should be selected by jointly considering fraud capture (recall) and false-alarm control (precision), with F1-score providing a useful summary of the balance between these priorities. Despite strong SMOTE-based performance,

some fraud cases remained undetected, indicating room for improvement. Future work should explore methods to further reduce false negatives while maintaining acceptable false-positive rates, including hybrid resampling strategies, threshold optimization, cost-sensitive learning, focal-loss training, probability calibration, and ensemble techniques. Evaluating robustness across multiple random splits and decision thresholds would also strengthen evidence for generalization and deployment readiness.

Declarations

Author Contributions: Conceptualization, Abdul Ahad Hassan Farroqi, Shah Ameer, Dr. Zahoor Ahmad, and Fahad Hassan Farooqi; methodology, Abdul Ahad Hassan Farroqi; software, Abdul Ahad Hassan Farroqi; validation, Abdul Ahad Hassan Farroqi, Shah Ameer, and Dr. Zahoor Ahmad; formal analysis, Abdul Ahad Hassan Farroqi; investigation, Abdul Ahad Hassan Farroqi; resources, Abdul Ahad Hassan Farroqi; data curation, Abdul Ahad Hassan Farroqi; writing—original draft preparation, Abdul Ahad Hassan Farroqi; writing—review and editing, Abdul Ahad Hassan Farroqi, Shah Ameer, Dr. Zahoor Ahmad, and Fahad Hassan Farooqi; visualization, Abdul Ahad Hassan Farroqi; supervision, Dr. Zahoor Ahmad; project administration, Abdul Ahad Hassan Farroqi. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: Publicly available datasets were analyzed in this study. The data can be found here: **Kaggle Credit Card Fraud Detection dataset** (<https://www.kaggle.com/mlg-ulb/creditcardfraud>).

Funding: Funding information is not available.

Acknowledgements: The authors would like to thank the Kaggle community and the dataset providers for making the credit card fraud detection dataset publicly available for research and educational purposes.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Conflicts of Interest: The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication

and/or submission, and redundancies have been completely observed by the authors.

References

- [1] CHANG V., ALI B., GOLIGHTLY L., GANATRA M. A., and MOHAMED M. Investigating credit card payment fraud with detection methods using advanced machine learning. *Information*, 2024, 15(8): 478. <https://doi.org/10.3390/info15080478>
- [2] ALBALAWI T., and DARDOURI S. Enhancing credit card fraud detection using traditional and deep learning models with class imbalance mitigation. *Frontiers in Artificial Intelligence*, 2025, 8: 1643292. <https://doi.org/10.3389/frai.2025.1643292>
- [3] WU Y., WANG L., LI H., and LIU J. A deep learning method of credit card fraud detection based on continuous-coupled neural networks. *Mathematics*, 2025, 13(5): 819. <https://doi.org/10.3390/math13050819>
- [4] AHMED M., ANSAR K., MUCKLEY C. B., KHAN A., ANJUM A., and TALHA M. A semantic rule based digital fraud detection. *PeerJ Computer Science*, 2021, 7: e649. <https://doi.org/10.7717/peerj-cs.649>
- [5] BTOUSH E. A. L. M., ZHOU X., GURURAJAN R., CHAN K. C., GENRICH R., and SANKARAN P. A systematic review of literature on credit card cyber fraud detection using machine and deep learning. *PeerJ Computer Science*, 2023, 9: e1278. <https://doi.org/10.7717/peerj-cs.1278>
- [6] RTAYLI N., and ENNEYA N. Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameter optimization. *Journal of Information Security and Applications*, 2020, 55: 102596. <https://doi.org/10.1016/j.jisa.2020.102596>
- [7] CHAWLA N. V., BOWYER K. W., HALL L. O., and KEGELMEYER W. P. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 2002, 16: 321–357. <https://doi.org/10.1613/jair.953>
- [8] HE H., and GARCIA E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 2009, 21(9): 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- [9] GUO H., LI Y., SHANG J., GU M., HUANG Y., and GONG B. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 2017, 73: 220–239. <https://doi.org/10.1016/j.eswa.2016.12.035>
- [10] HAFEZ I. Y., HAFEZ A. Y., SALEH A., ABD EL-MAGEED A. A., and ABOHANY A. A. A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of Big Data*, 2025, 12(1): 6. <https://doi.org/10.1186/s40537-024-01048-8>
- [11] ZIOVIRIS G., KOLOMVATSOS K., and STAMOULIS G. Credit card fraud detection using a deep learning multistage model. *The Journal of Supercomputing*, 2022, 78(12): 14571–14596. <https://doi.org/10.1007/s11227-022-04465-9>
- [12] FU K., CHENG D., TU Y., and ZHANG L. Credit card fraud detection using convolutional neural networks. In: *Neural Information Processing (ICONIP 2016)*, 2016, pp. 483–490. https://doi.org/10.1007/978-3-319-46675-0_53
- [13] KUMARI S. Enhanced deep neural network with SMOTE for credit card fraud detection. *EAI Endorsed Transactions*, 2025. <https://doi.org/10.4108/eai.28-4-2025.2357972>
- [14] DAVIS J., and GOADRICH M. The relationship between Precision–Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006, pp. 233–240. <https://doi.org/10.1145/1143844.1143874>
- [15] DAL POZZOLO A., CAELEN O., JOHNSON R. A., and BONTEMPI G. Calibrating probability with undersampling for unbalanced classification. In: *2015 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2015, pp. 159–166. <https://doi.org/10.1109/SSCI.2015.33>
- [16] FAWCETT T. An introduction to ROC analysis. *Pattern Recognition Letters*, 2006, 27(8): 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [17] SAITO T., and REHMSMEIER M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 2015, 10(3): e0118432. <https://doi.org/10.1371/journal.pone.0118432>

参考文献:

- [1] CHANG V., ALI B., GOLIGHTLY L., GANATRA M. A., MOHAMED M. 基于先进机器学习方法的信用卡支付欺诈检测研究. *Information*, 2024, 15(8): 478. <https://doi.org/10.3390/info15080478>
- [2] ALBALAWI T., DARDOURI S. 结合类别不平衡缓解的传统与深度学习信用卡欺诈检测研究. *Frontiers in Artificial Intelligence*, 2025, 8: 1643292. <https://doi.org/10.3389/frai.2025.1643292>
- [3] WU Y., WANG L., LI H., LIU J. 基于连续耦合神经网络的信用卡欺诈检测深度学习方法. *Mathematics*, 2025, 13(5): 819. <https://doi.org/10.3390/math13050819>
- [4] AHMED M., ANSAR K., MUCKLEY C. B., KHAN A., ANJUM A., TALHA M. 基于语义规则的数字欺诈检测方法. *PeerJ Computer Science*, 2021, 7: e649. <https://doi.org/10.7717/peerj-cs.649>
- [5] BTOUSH E. A. L. M., ZHOU X., GURURAJAN R., CHAN K. C., GENRICH R., SANKARAN P. 基于机器学习与深度学习的信用卡网络欺诈检测研究综述. *PeerJ Computer Science*, 2023, 9: e1278. <https://doi.org/10.7717/peerj-cs.1278>
- [6] RTAYLI N., ENNEYA N. 基于 SVM 递归特征消除与超参数优化的增强型信用卡欺诈检测. *Journal*

- of Information Security and Applications*, 2020, 55: 102596. <https://doi.org/10.1016/j.jisa.2020.102596>
- [7] CHAWLA N. V., BOWYER K. W., HALL L. O., KEGELMEYER W. P. SMOTE : 合成少数类过采样技术. *Journal of Artificial Intelligence Research*, 2002, 16: 321–357. <https://doi.org/10.1613/jair.953>
- [8] HE H., GARCIA E. A. 不平衡数据学习研究. *IEEE Transactions on Knowledge and Data Engineering*, 2009, 21(9): 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>
- [9] GUO H., LI Y., SHANG J., GU M., HUANG Y., GONG B. 类别不平衡数据学习: 方法与应用综述. *Expert Systems with Applications*, 2017, 73: 220–239. <https://doi.org/10.1016/j.eswa.2016.12.035>
- [10] HAFEZ I. Y., HAFEZ A. Y., SALEH A., ABD EL-MAGEED A. A., ABOHANY A. A. 信用卡欺诈检测中 AI 增强技术的系统综述. *Journal of Big Data*, 2025, 12(1): 6. <https://doi.org/10.1186/s40537-024-01048-8>
- [11] ZIOVIRIS G., KOLOMVATSOS K., STAMOULIS G. 基于深度学习多阶段模型的信用卡欺诈检测. *The Journal of Supercomputing*, 2022, 78(12): 14571–14596. <https://doi.org/10.1007/s11227-022-04465-9>
- [12] FU K., CHENG D., TU Y., ZHANG L. 使用卷积神经网络进行信用卡欺诈检测. In: *Neural Information Processing (ICONIP 2016)*, 2016, pp. 483–490. https://doi.org/10.1007/978-3-319-46675-0_53
- [13] KUMARI S. 基于 SMOTE 的增强型深度神经网络信用卡欺诈检测. *EAI Endorsed Transactions*, 2025. <https://doi.org/10.4108/eai.28-4-2025.2357972>
- [14] DAVIS J., GOADRICH M. Precision–Recall 曲线与 ROC 曲线的关系. In: *ICML 会议论文集*, 2006, pp. 233–240. <https://doi.org/10.1145/1143844.1143874>
- [15] DAL POZZOLO A., CAELEN O., JOHNSON R. A., BONTEMPI G. 不平衡分类中结合欠采样的概率校准方法. In: *IEEE SSCI 会议论文集*, 2015, pp. 159–166. <https://doi.org/10.1109/SSCI.2015.33>
- [16] FAWCETT T. ROC 分析导论. *Pattern Recognition Letters*, 2006, 27(8): 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- [17] SAITO T., REHMSMEIER M. 在不平衡数据上评估二分类器时, Precision–Recall 曲线比 ROC 曲线更具信息量. *PLOS ONE*, 2015, 10(3): e0118432. <https://doi.org/10.1371/journal.pone.0118432>

Manuscript Information

Word count: 7,213 words (excluding references).

Peer-Review Record

Fast-track status: Not fast-tracked.

First-round reviews received: 3 reports.

Revision cycles completed: 3 rounds.

Final version submitted: April 6, 2026

Disclaimer / Publisher's Note

The statements, opinions, and data contained in this article are solely those of the authors and do not necessarily represent the views of the *Journal of Hunan University (Natural Sciences)* or its editorial team. The journal and its editors disclaim any responsibility for injury to persons or property resulting from any ideas, methods, instructions, or products referred to in the content of this article.