*International Journal of Computational and Experimental Science and ENgineering (IJCESEN)*

*Vol. 11-No.3 (2025) pp. 5153-5163*
*http://www.ijcesen.com*

***ISSN: 2149-9144***



**Research Article**

# Improving Imbalanced Data Classification Using Deep Learning

**Nihaya S. Salih[1], Dindar M. Ahmed[2*]**

[1] Department of Information Technology, Technical College of Duhok, Duhok Polytechnic University, Duhok, Kurdistan Region–Iraq
**Email:** nihaya.salih@dpu.edu.krd - **ORCID**:0000-0002-1321-0220

[2] Department of Information Technology Management, Technical College of Administration, Duhok Polytechnic University, Duhok, Kurdistan Region–Iraq,
* **Corresponding Author Email:** dindar.ahmed@dpu.edu.krd- **ORCID:** 0000-0002-1321-0221

**Article Info:**

**DOI:** 10.22399/ijcesen.3367
**Received :** 22 May 2025
**Accepted :** 06 July 2025

**Keywords**

Imbalanced Data
Fraud Detection
Deep Learning
Neural Networks
Ensemble Learning

**Abstract:**

Classifying imbalanced data is a difficult task in many machine learning applications, especially in the context of fraud detection. This paper evaluated the performance of traditional models (e.g., Random Forests, XGBoost, and CatBoost) against the performance of deep learning models. While the traditional models were able to obtain high accuracy, they struggled to identify the rare classes (i.e., fraudulent transactions) when the F1 scores did not get above 0.33. In turn, a deep learning model was proposed that applied ideas such as class weights, decision thresholds, and F1-maximizing training objectives and was designed to employ voting of multiple submodels. The results demonstrated that the proposed model (Ensemble Neural Network) was able to achieve an F1 score of 0.5997 and an AUC-PR score of 0.6205 which outperformed the traditional methods previously used in the study. This design was used to achieve a better balance between identifying the rare classes and overall model performance.

## 1. Introduction

Machine learning (ML) has rapidly become part of decision-making systems in numerous fields such as finance, healthcare, cybersecurity, and e-commerce in recent years [1,2]. While it has enjoyed considerable success in many applications, one persistent problem for machine learning algorithms is the problem of data imbalance, in which one class (often the class of interest) is severely underrepresented compared to a majority class [3,4]. This problem poses significant challenges in high-stakes applications, such as fraud detection, in which ignoring rare events (e.g., fraudulent transactions) can cause substantial loss of funds and/or reputation.[5,6] Standard machine learning models such as random forests, decision trees, and gradient boosting techniques (such as CatBoost and XGBoost) have been successful for many classification problems [7,8]. However, their performance degrades dramatically with datasets with very imbalanced classes [9]. This is due to the fact that overall accuracy is often our modeling goal, which essentially favors the majority class and hinders the detection of minority instances [10]. There have been numerous methods proposed in the literature that have been designed to develop a better class model, ranging from data-level techniques (such as SMOTE and under-sampling) to algorithmic techniques (such as cost-sensitive learning), to threshold tuning. SMOTE and other synthetic oversampling techniques have been especially useful as they allow the generation of realistic samples of the minority classes [11,12]. However, both SMOTE and other oversampling techniques are generally insufficient to overcome under-sampling with real-world problems that contain large amounts of complex, high-dimensional data. Simultaneously, deep learning (DL) is an alternative that also has the potential to learn complicated, nonlinear relationships and capture nuanced patterns in large-scale data.[13,14] Importantly, deep neural networks offer an advantage of flexibility in the use of advanced training techniques that were developed specifically for addressing class imbalance, such as class-weighted loss functions, custom evaluation metrics, dynamic thresholds, and group-based structures[15,16]. In recent work, it was shown that these models were able to improve recall and

accuracy for minority classes without major detriment to overall performance [17]. Machine Learning (ML) is a discipline of artificial intelligence (AI) that allows computer systems to learn from data, improving their performance without being explicitly programmed or needing to have their explicit rules refreshed [18]. Just as humans learn from experience, the system identifies patterns, relationships, or rules by analyzing large volumes of analyzed and historical data in what is called "supervised learning"[19]. When presented with new, unobserved inputs, these systems can make predictions or decisions concerning these new observations [20]. ML algorithms can be classified into three general types: supervised learning, unsupervised learning, and reinforcement learning [21] . In the case of classification problems such as fraud detection, supervised learning is important, where labelled datasets are used to build and fit a model with actual examples that separate two classes (i.e., fraudulent transactions and non-fraudulent transactions) [22]. ML models in classification settings are designed to find a decision boundary with minimal error separating the classes of interest [23] . Common algorithms for those tasks include Decision Trees, Random Forests, Support Vector Machines (SVM), and Gradient Boosting methods like XGBoost and CatBoost[24]. Traditional ML models have a serious limitation in that they do not function well if we have imbalanced datasets, i.e. all the observations represent a majority class with some observations representing a minority class [25] . In imbalanced datasets, the model tends to favor the majority class, resulting in poor performance for the minority class, which is typically the class of greatest interest in applications such as fraud detection and rare diseases [26] . Deep Learning (DL) is a distinct type of machine learning that involves neural networks with several layers (deep architectures) to model complex, nonlinear relationships in data [27]. Deep neural networks (DNNs) are inspired by the human brain; consisting of multiple layers of neurons, which are interconnected, are transformed by weighted connections that perform an activation function [27]. DNNs are immensely effective for learning from high-dimensional and unstructured data like images, text, and time-series data.[28] When working with classification problems, deep learning (DL) models are trained with backpropagation to minimize a loss function—binary cross-entropy, for instance, can be used for binary classification [29] . One of the main benefits of DL models is their ability to automatically learn hierarchical features from raw data; no feature engineering is required [30]. Architechture configuration, loss function and training customization options are also plentiful with DL frameworks, making them adaptable to complex tasks, such as classification of imbalanced data [31]. To mitigate the class imbalance implications, advanced methodologies in deep learning include - class-weighted loss functions, customized thresholds, early stopping, and ensemble methods to enable continued training with consideration of the majority and minority classes and improve generalization [32]. Notably, if the model penalizes the minority during the training more heavily than the majority through the loss function, for example, the higher-class weight for the minority, then the network will consider it more during its decision-making, ultimately making a better decision. Similarly, thresholding with the probability score can be lower for the minority class than the majority class [33,34]. Imbalanced data refers to a classification problem in which the classes are not represented equally. Typically, one class (the majority class) has a significantly higher number of instances than the other (the minority class). [35] This imbalance poses a serious challenge for machine learning algorithms, as most standard models are designed to maximize overall accuracy, leading to a bias toward predicting the majority class correctly while ignoring the minority class [36]. In many real-world applications—such as fraud detection, medical diagnosis, and fault detection, the minority class represents critical events that are far more important to detect accurately despite their rarity [37]. The main issue with imbalanced datasets is that conventional evaluation metrics like accuracy become misleading [38]. A model can achieve high accuracy simply by predicting the majority class every time, while completely failing to identify the minority class.[39] To properly evaluate model performance on imbalanced data, alternative metrics such as precision, recall, F1-score, and the area under the precision-recall curve (AUC-PR) are preferred, as they emphasize the model's ability to identify rare but significant instances[40]. There are a variety of specialized methods that can be used when dealing with class imbalance - both data-level techniques (for example, by oversampling the minority class, such as through SMOTE, undersampling the majority class, or the combination of the two...) and algorithm-level methods (for example: altering the loss function of the model [e.g. cost-sensitive learning], rebalancing the decision threshold, and using ensembles that emphasize the performance of the minority classes)[41]. In the context of deep learning, we can handle imbalance by using approaches such as class-weight learning, tuning threshold, using custom loss for evaluation, etc. The effectiveness of

these approaches will depend highly upon the properties of the dataset and the domain in which they are conveyed [42].

This study aims to assess and compare performance of traditional ensemble classifiers and custom deep learning frameworks on fraud detection tasks. In this study's proposed deep learning model, there is a unique combination of class weighting, threshold adjustment, and F1 score optimization that result in an accumulation of the results through an electronic voting of ensembles.

## 2. Related Work

Many approaches have been proposed to handle class imbalance, including resampling techniques, cost-sensitive learning, and ensemble methods. Deep learning has emerged as a powerful tool capable of capturing complex patterns in data. Previous studies have shown that incorporating loss function modifications, batch normalization, early stopping, and ensemble strategies can significantly enhance model performance on imbalanced datasets. Khatir et al. [43] examined machine learning for credit scoring. They used feature selection and oversampling techniques to compare five classifiers. The German credit dataset was used for the experiments. Random Forest with RFE and Random Oversampling produced the best results. Akinjole et al. [44] A comprehensive ensemble framework integrating classical ML and deep learning has been created for credit default prediction. The models included XGBoost, RF, DT, SVM, ADABoost, and a three-layer MLP. The approach utilized RFECV for feature selection and SMOTE+ENN for balancing. The stacking ensemble scored 93.69% accuracy and 0.9781 AUC, indicating good predicting performance. Alagić et al. [45] A comparative machine learning framework was constructed to improve credit risk prediction utilizing both financial and mental health data. Two datasets were trained using algorithms such as XGBoost, RF, DT, KNN, AdaBoost, and GBoost. Following preprocessing and balanced splitting, XGBoost and RF demonstrated the highest accuracy, precision, recall, and F1-score. Chaturvedi et al. [46] Using a Kaggle dataset, the researchers tested several machine learning models for forecasting non-performing loans. Models used were RF, XGBoost, GBoost, LSTM, DT, Naïve Bayes, and LightGBM. Preprocessing was done using SMOTE, normalization, and one-encoding. According to criteria such as AUC-PR and F1-score, Random Forest performed the best on unbalanced data. Zhao et al. [47] Authers used four datasets to evaluate resampling approaches for credit risk prediction. Nine approaches, including SMOTE, ENN, and SH-SENN, were examined with 11 classifiers via cross-validation and Bayesian optimization. SH-SENN, a combination of SMOTE and ENN, performed best on highly unbalanced data. Yang and Xiao [48] A multi-stage ensemble approach for assessing SME credit risk was presented, integrating financial and soft features. The method utilized bagging-based SMOTE for balancing and L-BFGS-B for adaptive voting optimization. This strategy improved both interpretability and classification robustness. Abidemi et al. [49] This strategy improved both interpretability and classification robustness. They used PCA, normalizing, and dummy encoding to engineer features. Imbalanced data was treated with oversampling and undersampling approaches. Model performance was assessed using accuracy, precision, recall, F1-score, and AUC-ROC. Long et al. [50] proposed a credit scoring model that combines financial literacy elements and ensemble machine learning methods. Standardization, SMOTE, and chi-square feature selection were among the preprocessing steps. Models employed included RF, AdaBoost, GBDT, LightGBM, and Voting Classifier. Accuracy, F1-score, and AUC across multiple feature sets were used to gauge performance. Paudel et al. [51] Created a multi-class credit risk prediction model with a Kaggle dataset and deep learning architectures, notably GRU and Bi-LSTM. They used SMOTE-ENN for class balance, weighted F1-score for performance evaluation across different train-test splits, correlation and information value measures for feature selection, and thorough preprocessing. Zhuang et al. [52] A credit risk model that combined DNN and the Improved Butterfly Optimization Algorithm (IBOA) was presented in the study. RFE was used for feature selection, while SMOTE was used to address class imbalance. To improve performance, DNN parameters were optimized using IBOA. On benchmark datasets, the IBOA-DNN performed better than conventional ML models. Liang et al. [53] Combined CNN with SHAP-based feature weighting to create an interpretable credit rating model. For training, input characteristics were ranked and reweighed using SHAP values. Credit datasets from Australia and Germany were used to test the model. It maintained excellent deep learning performance while enhancing transparency. Yang et al. [54] The Authors suggested an ensemble neural network approach for predicting loan default. The training data was balanced using random undersampling and SMOTE.

Ensemble averaging was used to collect the results from base classifiers. Precision, recall, and AUC

were used to validate the model on data from Lending Club.

## 4. Methodology

### 4.1 Dataset Description

The data from Kaggle labeled "Credit Card Fraud Detection" was utilized. It contains credit card transactions from European customers of a bank over a two-day period in September 2013. There are a total of 284,807 transactions, of which only 492 were 'fraudulent.' This is very biased toward class imbalance, since 'fraudulent' transactions were only about 0.172% of the complete data. Fig 1 shows the distribution of classes for the data set.


***Figure 1.*** *shows the distribution of categories representing 0 no fraud and 1 credit fraud.*

### 4.2 Data Preprocessing

In order to methodically prepare the financial transaction data for fraud detection, a sequence of preprocessing operations were conducted. First, the "Unnamed: 0" column corresponding to the autosave process was deleted, and columns with maximum correlation or no immediate analytical value such as zip, unix_time, merch_lat, and lat were all removed at this time to eliminate redundancy and improve model performance. Then, the date column of type string (trans_date_trans_time) was converted to the datetime type and individual features extracted as strings from it, including year, month, day, hour, minute, which were useful to capture the temporal aspects of fraudulent behavior. The original column was now deleted. The categorical variables were automatically detected using data type and were also encoded numerically using LabelEncoder. The training and test set were combined briefly to ensure consistent encoding. Then the StandardScaler was used for the numerical features to bring values to a mean of 0 and standard deviation of 1; making the training more stable which further improve model performance, especially deep learning networks. At a later date, the target variable (is_fraud) was split away from former features and created sets (X_train, y_train) for training, and (X_val, y_val) for evaluation. A numerical correlation matrix was created and visualized with a displayed heat map in order to see which variables overlap.

### 4.3 Proposed Model

The current research develops a framework based on deep learning to solve the problem of imbalance classification in fraud detection. The framework has three necessary parts; data preprocessing, special neural network models, and an ensemble decision-making process. The data preprocessing approach required the dataset to be normalized and stratified to preserve the distribution of class labels. The associated models included three different deep neural network architectures, each of which was designed to combat imbalance classification in different ways. Model 1 introduces a weighting factor into the loss function so that misclassifications occur more heavily regarding the minority class. Model 2 introduced thresholds as an approach by lowering the threshold below a score of 0.5 to improve the model's ability to classify rare occurrences. Model 3 optimized for F1 and used ad hoc training calls and early stopping so the results would maximize the geometric mean of precision and recall. The uncertain predictions from each model were then averaged together through an ensemble voting method. This ensemble voting method enables the predictions from the three models to create a level of confidence for decision making and creates a robust model that generalizes better. The final output from the overall framework

was a modified neural network model that is able to achieve greater performance in detecting the minority classes, and achieved better F1 and AUC-PR ratios than previous methods. This architecture took advantage of model diversity and is able to create a relevant, fast end-user solution for real-world imbalanced data issues. Figure 2 illustrates the stages of the proposed methodology.

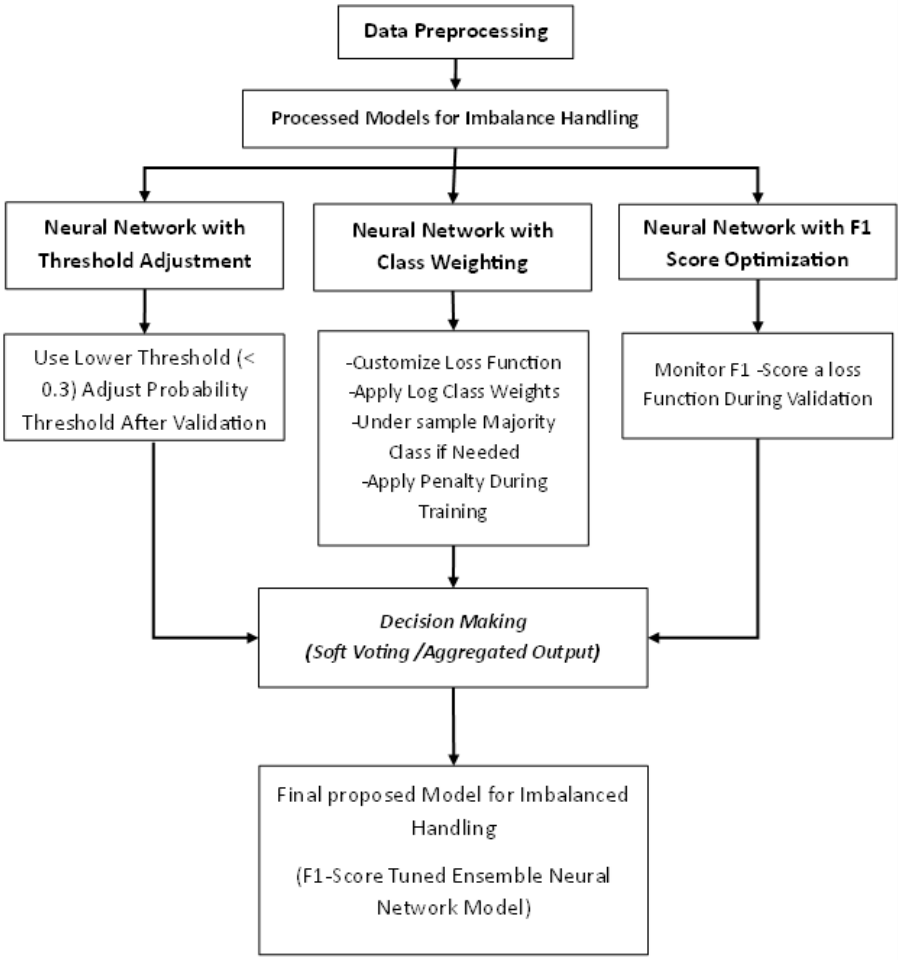


***Figure 2.*** *Illustrates the Stages of The Proposed Methodology*

**Model 1: Class Weighting**

This model changes the standard binary cross-entropy loss function by adding class-based weights. We have provided a higher weight to the minority class. This means that during training we will penalize the misclassification more heavily. Importantly, the weighted loss function will allow the model to devote more learning capacity to the under-represented instances. The weighted loss function is defined as:

$$L_i = -\sum_{i=1}^{N} w_{y_i} [y_i \log P_i + (1 - y_i) \log(1 - p_i)] \quad \text{……..(1)}$$

where:

- $p_i = f_1(x_i)$ : The predicted probability of the first model.
- $w_{y_i}$ : The class weight, defined as:

$$w_{y_i} = \begin{cases} w_0 & \text{if } y_i = 0 \\ w_1 & \text{if } y_i = 1, w_1 > w_0 \end{cases} \quad \text{……..(2)}$$

**Model 2: Threshold Adjustment**

This model replaces standard probability thresholds in binary classification (generally 0.5) with an empirically determined threshold that is lower. This encourages the model to classify borderline cases as positive, increasing recall for the minority class. The tuning of the threshold uses validation set performance. The final prediction is made with a tunable threshold. $\tau \in (0,1)$

$$\hat{y}_i^{(2)} = \begin{cases} 1 & \text{if } f_2(x_i) \geq \tau \\ 0 & \text{otherwise} \end{cases} \quad \text{……..(3)}$$

Where:

$f_2(x_i)$ is the probability prediction for class 1 from the second model/

**Model 3: F1-Score Optimization**

Rather than optimizing for accuracy, this model relies on a custom training callback that measures the F1-score on the validation set. Training is directed to maximize the harmonic mean of precision and recall, and early stopping is applied at a plateau of the F1-score. This approach is specifically intended to maximize balanced performance on a skewed class distribution. The F1-Score is calculated as:

$$F1 - Score = 2.\frac{Precision.Recall}{Precision+Recall} \quad \text{.....(4)}$$

This score is used as an indirect training objective via a callback , where the loss is :

$$L_3 = 1 - F1_{val} \quad \text{.....(5)}$$

Early stopping is applied when $F1_{val}$ reaches its maximum.

**Ensemble Voting Strategy**

After training the three models separately, we combined their outputs in a soft-ensemble voting fashion. Each of the models generates a probability score for the positive class, and the final prediction is then calculated as the average of the three probabilities:

$$P_{final^{(x)}} = \frac{1}{3}\left[f_{1^{(x)}} + f_{2^{(x)}} + f_{3^{(x)}}\right] \quad \text{.....(6)}$$

The Final class prediction is made as :

$$\hat{y}_{\text{final}} = \begin{cases} 1 & \text{if } P_{\text{final}}(x) \geq \theta \\ 0 & \text{otherwise} \end{cases} \quad \text{.....(7)}$$

where $\theta$ is the decision threshold used in the ensemble model.

**4.4 Evaluation Metrics**

In imbalanced classification tasks such as fraud detection, it can be misleading to rely solely on traditional metrics such as accuracy[55,56] A model that is capable of predicting almost all transactions as non-fraudulent experiences high accuracy, then detects no actual fraud. Consequently, we discuss metrics that are more fitting for evaluating the model's performance in recognizing the minority class.[57]

**Precision**

Precision tells us how many of the predicted fraud cases were actually correct:

$$\text{Precision} = \frac{TP}{TP+FP} \quad \text{..... (8) [58]}$$

Where:

- **TP** = True Positives (correct fraud predictions)
- **FP** = False Positives (incorrect fraud predictions)

A high precision means fewer false alarms, which is valuable in real-world scenarios where each false alert can be costly.

**Recall (Sensitivity or True Positive Rate)**

Recall measures how many of the actual fraud cases were correctly identified:

$$\text{Recall} = \frac{TP}{TP+FN} \quad \text{.....(9) [59]}$$

- **FN** = False Negatives (missed fraud cases)

In highly imbalanced datasets, recall is critical—it shows how well the model catches rare events like fraud. A low recall means many fraudulent cases go undetected.

**F1-Score**

The F1-score balances precision and recall by calculating their harmonic mean:

$$\text{F1-Score} = 2.\frac{Precision.Recall}{Precision+Recall} \quad \text{.....(10) [60]}$$

This score is especially useful when both false positives and false negatives are costly—like accusing a customer wrongly or missing actual fraud.

**Accuracy**

Accuracy measures the overall proportion of correct predictions:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{.....(11) [60]}$$

While it's reported for completeness, it is not a reliable metric in imbalanced scenarios, since it can be high even when the model misses most fraud cases.

**Area Under the Precision-Recall Curve (AUC-PR)**

AUC-PR summarizes the precision-recall trade-offs for all classification thresholds. It is more meaningful than AUC-ROC for imbalanced data, as it only considers the performance of the minority class. A higher AUC-PR indicates that the model ranks actual frauds higher than non-frauds[61].

## 5. Experiments and Results

To compare various strategies for dealing with imbalanced data for fraud detection, multiple empirical experiments were conducted using the Credit Card Fraud Detection dataset. Empirical experiments were implemented in Python 3.11, using: scikit-learn for machine learning models; XGBoost and CatBoost for ensemble methods; TensorFlow/Keras for deep learning; and Matplotlib and Seaborn for visualization. All models were trained on a provided training set and then tested on a separate test set (stratified sampling was used to maintain the proportions of original class distribution). Table 1 gives the performance of traditional machine learning models, built with no balancing adjustments. These models were our baseline to compare subsequent results.

***Table 1.*** *Traditional Machine Learning Models*

| Model | Precision | Recall | F1-Score | AUC-PR | Accuracy |
|---|---|---|---|---|---|
| Random Forest | 0.2882 | 0.3879 | 0.3288 | 0.3092 | 0.9845 |
| XGBoost | 0.1135 | 0.2485 | 0.1471 | 0.2092 | 0.9733 |
| CatBoost | 0.0866 | 0.2747 | 0.1390 | 0.3041 | 0.9641 |

Table 2 shows results for baseline neural networks, including variants with class weighting and oversampling. All models were designed to further improve overall sensitivity to the minority class by adjusting either training inputs or loss function penalties.

***Table 2.*** Baseline and Oversampling Neural Networks

| Model | Precision | Recall | F1-Score | AUC-PR | Accuracy |
|---|---|---|---|---|---|
| Baseline NN | 0.0688 | 0.6182 | 0.1238 | 0.4680 | 0.9134 |
| NN + class_weight | 0.0688 | 0.6182 | 0.1238 | 0.4680 | 0.9134 |
| NN + class_weight + Oversampling | 0.1050 | 0.5697 | 0.1774 | 0.4002 | 0.9477 |

Table 3 shows results from more advanced deep learning models that used training-level optimization techniques, including early stopping, batch normalization, F1-score stimulus, and ensemble voting, to attempt to improve upon fraudulent case identification.

***Table 3.*** Optimized Deep Learning Models

| Model | Precision | Recall | F1-Score | AUC-PR | Accuracy |
|---|---|---|---|---|---|
| NN + EarlyStopping + BN | 0.7320 | 0.3283 | 0.4533 | 0.4508 | 0.9922 |
| NN + F1-Callback | 0.6836 | 0.3667 | 0.4773 | 0.4489 | 0.9921 |
| NN + Ensemble Voting | 0.6406 | 0.5636 | 0.5997 | 0.6205 | 0.9925 |

## 6. Discussion

This study's results highlighted that resolving data imbalance is not merely a technical choice, but a clear requirement, especially in important spaces, such as financial fraud detection. When using traditional models like Random Forest, XGBoost, and CatBoost mean accuracy rates were high overall, but their predictions did not yield similarly high accuracy rates for fraudulent values, which are integral to the task. For example, while Random Forest had an overall accuracy of over 98%, it had a poor response to minority classes as demonstrated by its low F1-Score and low AUC-PR values.

On the other hand, both when using deep neural network models, and combining deep with imbalance-correcting techniques, such as class weights, threshold adjustment, and correcting the loss function for F1-Score, we highlighted the clear improvement in performance which was not just an increase in numerical indicators but implicit in the number of frauds discovered and false alarms decreased. A particularly good example of this was in our proposed model. This model used cumulative

voting of three-5 networks, and the final models were the best in terms of revisiting limitations regarding precision and recall, resulting in an F1-Score of 0.5997 and AUC-PR of 0.6205; of which no models prior, were matched.

The unique aspect of the model is its deep architecture, and combined with proposed model of three networks, which addressed three different aspects of the imbalance problem. One relies on one model weighted classes to give more importance to rare cases, one reduces the prediction threshold, and the final one tracks improvement in the F1-Score. We combined the results using the soft-voting method which can use more balanced methods and often allows more flexibility in predicting options.

It is therefore the case that it is not sufficient for the models to only be trained on a powerful model; the training procedure must be well thought out, it is necessary to know what is consistent with the data, and performance should be a reflection of achieving a model's goal measured with sensitive metrics that motivate the model behaviour. The study has provided evidence that when deep models are systematically trained and are well posed, they can yield accurate results in the most challenging circumstances. There are rich possibilities for their continued use in detecting fraud and in other contexts characterized by sensitive validities.

## 7. Conclusion

The study demonstrates that tackling imbalanced data involves more than building a strong classification model. The study concludes that traditional models can achieve high accuracy when prediction is based on the data as a whole, but they are unsuitable for predicting rare classes like financial fraud. Through its methodology, the deep learning models built in this study were able to outperform these traditional and basic deep learning models by employing weighting the underrepresented class, adjusting the probabilistic threshold, and optimizing the loss function based on the F1-Score. Three models employing these methodologies generated its final model using an ensemble voting scheme. It achieved a favourable stored balance of fraud detection to false alarm, and outperformed all other models in precision, recall, F1-Score and AUC-PR. Thus, neural network models augmented with form of training that take into account imbalance are acceptable in sensitive areas like fraud detection where errors are costly. The study highlights that there is evidence that the kind of model, as well as different kinds of models combined into an intelligent voting scheme can generate more robust models for predictions.

## Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## References

[1] Razzaq, K., & Shah, M. (2025). Machine Learning and Deep Learning Paradigms: From Techniques to Practical Applications and Research Frontiers. *Computers*, *14*(3), 93.

[2] Dritsas, E., & Trigka, M. (2025). Exploring the Intersection of Machine Learning and Big Data: A Survey. *Machine Learning and Knowledge Extraction*, *7*(1), 13.

[3] Ghosh, K., Bellinger, C., Corizzo, R. *et al.* The class imbalance problem in deep learning. *Mach Learn* **113**, 4845–4901 (2024).

[4] Altalhan, Manahel & Algarni, Abdulmohsen & Monia, Turki. (2025). Imbalanced Data Problem in Machine Learning: A Review. IEEE Access. PP. 1-1. 10.1109/ACCESS.2025.3531662.

[5] Theodorakopoulos, L., Theodoropoulou, A., Tsimakis, A., & Halkiopoulos, C. (2025). Big Data-Driven Distributed Machine Learning for Scalable Credit Card Fraud Detection Using PySpark, XGBoost, and CatBoost. *Electronics*, *14*(9), 1754.

[6] Kim, H. (2025). Novel Deep Learning-Based Facial Forgery Detection for Effective Biometric Recognition. *Applied Sciences*, *15*(7), 3613.

[7] Kopyt, M., Piotrowski, P., & Baczyński, D. (2024). Short-Term Energy Generation Forecasts at a Wind Farm—A Multi-Variant Comparison of the Effectiveness and Performance of Various Gradient-Boosted Decision Tree Models. *Energies*, *17*(23), 6194.

[8] Kumar, V., Kedam, N., Sharma, K. V., Khedher, K. M., & Alluqmani, A. E. (2023). A Comparison of Machine Learning Models for Predicting Rainfall in Urban Metropolitan Cities. *Sustainability*, *15*(18), 13724.

[9] Aguilar-Ruiz, J.S., Michalak, M. Classification performance assessment for imbalanced multiclass data. *Sci Rep* **14**, 10759 (2024).

[10] Pagano, T. P., Loureiro, R. B., Lisboa, F. V. N., Peixoto, R. M., Guimarães, G. A. S., Cruz, G. O. R., Araujo, M. M., Santos, L. L., Cruz, M. A. S., Oliveira, E. L. S., Winkler, I., & Nascimento, E. G. S. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. *Big Data and Cognitive Computing*, *7*(1), 15.

[11] Sakri, S., & Basheer, S. (2023). Fusion Model for Classification Performance Optimization in a Highly Imbalance Breast Cancer Dataset. *Electronics*, *12*(5), 1168.

[12] Yang, Yuxuan & Khorshidi, Hadi & Aickelin, Uwe. (2024). A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. Frontiers in Digital Health. 6. 1430245. 10.3389/fdgth.2024.1430245.

[13] Irfan, Muhammad & Mushtaq, Zohaib & Khan, Nabeel & Mursal, Salim & Rahman, Saifur & Magzoub, Muawia & Latif, Muhammad Armghan & Althobiani, Faisal & Khan Yousufzai, Imran & Abbas, Ghulam. (2023). A Scalo gram-based CNN Ensemble Method with Density-Aware SMOTE Oversampling for Improving Bearing Fault Diagnosis. IEEE Access. PP. 1-1. 10.1109/ACCESS.2023.3332243.

[14] Mazdadi, Muhammad & Saragih, Triando Hamonangan & Budiman, Irwan & Farmadi, Andi & Tajali, Ahmad. (2024). The Effectiveness of Data Imputations on Myocardial Infarction Complication Classification Using Machine Learning Approach with Hyperparameter Tuning. Jurnal Ilmiah Teknik Elektro Komputer dan Informatika. 10. 520-533. 10.26555/jiteki.v10i3.29479.

[15] Farhadpour, Sarah & Warner, Timothy & Maxwell, Aaron. (2024). Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices. Remote Sensing. 16. 533. 10.3390/rs16030533.

[16] Imrana, Y., Xiang, Y., Ali, L. *et al*. CNN-GRU-FF: a double-layer feature fusion-based network intrusion detection system using convolutional neural network and gated recurrent units. *Complex Intell. Syst.* **10**, 3353–3370 (2024).

[17] Palak Gupta, Anmol Varshney, Mohammad Rafeek Khan, Rafeeq Ahmed, Mohammed Shuaib, Shadab Alam, Unbalanced Credit Card Fraud Detection Data: A Machine Learning-Oriented Comparative Study of Balancing Techniques, Procedia Computer Science, Volume 218, 2023, Pages 2575-2584, ISSN 1877-0509.

[18] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN COMPUT. SCI.* **2**, 160 (2021).

[29] Razzaq, K., & Shah, M. (2025). Machine Learning and Deep Learning Paradigms: From Techniques to Practical Applications and Research Frontiers. *Computers*, *14*(3), 93.

[20] S. Cheng *et al.*, "Machine Learning With Data Assimilation and Uncertainty Quantification for Dynamical Systems: A Review," in *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 6, pp. 1361-1387, June 2023, doi: 10.1109/JAS.2023.123537.

[21] Eduardo F. Morales, Hugo Jair Escalante, Chapter 6 - A brief introduction to supervised, unsupervised, and reinforcement learning, Editor(s): Alejandro A. Torres-García, Carlos A. Reyes-García, Luis Villaseñor-Pineda, Omar Mendoza-Montoya, Biosignal Processing and Classification Using Computational Learning and Intelligence, Academic Press, 2022, Pages 111-129, ISBN 9780128201251.

[22] Ali, A., Abd Razak, S., Othman, S. H., Eisa, T. A. E., Al-Dhaqm, A., Nasser, M., Elhassan, T., Elshafie, H., & Saif, A. (2022). Financial Fraud Detection Based on Machine Learning: A Systematic Literature Review. *Applied Sciences*, *12*(19), 9637.

[23] Fazil, A. W., Hakimi, M., Akbari, R., Quchi, M. M., & Khaliqyar, K. Q. (2023). Comparative analysis of machine learning models for data classification: An in-depth exploration. *Journal of Computer Science and Technology Studies*, *5*(4), 160-168.

[24] Lijie Zhang, Dominik Jánošík, Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches, Expert Systems with Applications, Volume 241, 2024, 122686, ISSN 0957-4174.

[25] Sharma, S., & Gosain, A. (2025). Addressing class imbalance in remote sensing using deep learning approaches: a systematic literature review. *Evolutionary Intelligence*, *18*(1), 1-28. ISO 690

[26] M. Altalhan, A. Algarni and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," in *IEEE Access*, vol. 13, pp. 13686-13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[27] Taye, M. M. (2023). Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions. *Computers*, *12*(5), 91. https://doi.org/10.3390/computers12050091.

[28] Sakib, M., Mustajab, S., & Alam, M. (2025). Ensemble deep learning techniques for time series analysis: a comprehensive review, applications, open issues, challenges, and future directions. *Cluster Computing*, *28*(1), 1-44.

[29] Ruby, U., & Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. *Int. J. Adv. Trends Comput. Sci. Eng*, *9*(10). ISO 690

[30] Ahmed, S. F., Alam, M. S. B., Hassan, M., Rozbu, M. R., Ishtiak, T., Rafa, N., ... & Gandomi, A. H. (2023). Deep learning modelling techniques: current progress, applications, advantages, and challenges. *Artificial Intelligence Review*, *56*(11), 13521-13617. ISO 690

[31] Mienye, I. D., & Swart, T. G. (2024). A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information*, *15*(12), 755.

[32] M. Altalhan, A. Algarni and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," in *IEEE Access*, vol. 13, pp. 13686-13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[33] Fang, C., He, H., Long, Q., & Su, W. J. (2021). Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, *118*(43), e2103091118.

[34] Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: adjusting the decision threshold to handle imbalanced data in machine learning. *Journal of Chemical Information and Modeling*, *61*(6), 2623-2640.

[35] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, Amanda Gonsalves, Data imbalance in classification: Experimental evaluation, Information Sciences, Volume 513, 2020, Pages 429-441, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.11.004

[36] M. Altalhan, A. Algarni and M. Turki-Hadj Alouane, "Imbalanced Data Problem in Machine Learning: A Review," in *IEEE Access*, vol. 13, pp. 13686-13699, 2025, doi: 10.1109/ACCESS.2025.3531662.

[37] Calabrese, F., Regattieri, A., Bortolini, M., & Galizia, F. G. (2022). Data-driven fault detection and diagnosis: Challenges and opportunities in real-world scenarios. *Applied Sciences*, *12*(18), 9212.

[38] Owusu-Adjei, M., Ben Hayfron-Acquah, J., Frimpong, T., & Abdul-Salaam, G. (2023). Imbalanced class distribution and performance evaluation metrics: A systematic review of prediction accuracy for determining model performance in healthcare systems. *PLOS Digital Health*, *2*(11), e0000290.

[39] Douzas, G., Bacao, F., Fonseca, J., & Khudinyan, M. (2019). Imbalanced learning in land cover classification: Improving minority classes' prediction accuracy using the geometric SMOTE algorithm. *Remote Sensing*, *11*(24), 3040.

[40] Asare, M. (2024). *Evaluating Feature Selection Methods in Machine Learning With Class Imbalance* (Master's thesis, The University of Texas Rio Grande Valley).

[41] Wongvorachan, T., He, S., & Bulut, O. (2023). A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining. *Information*, *14*(1), 54.

[42] Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. *Machine Learning*, *113*(7), 4845-4901.

[43] Khatir, Ahmed & Bee, Marco. (2022). Machine Learning Models and Data-Balancing Techniques for Credit Scoring: What Is the Best Combination?. Risks. 10. 10.3390/risks10090169.

[44] Akinjole, Abisola & Shobayo, Olamilekan & Popoola, Jumoke & Okoyeigbo, Obinna & Ogunleye, Bayode. (2024). Ensemble-Based Machine Learning Algorithm for Loan Default Risk Prediction. Mathematics. 12. 3423. 10.3390/math12213423.

[45] Alagic, A., Zivic, N., Kadusic, E., Hamzic, D., Hadzajlic, N., Dizdarevic, M., & Selmanovic, E. (2024). Machine Learning for an Enhanced Credit Risk Analysis: A Comparative Study of Loan Approval Prediction Models Integrating Mental Health Data. *Machine Learning and Knowledge Extraction*, *6*(1), 53-

[46] T. Chaturvedi, S. Halder, U. S. kumar, N. Das and S. Bittu, "Comparative Performance Analysis of Machine Learning Algorithms for Non-Performing Loan Prediction," *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*, Indore, India, 2025, pp. 13-18, doi: 10.1109/ICCCIT62592.2025.10928008.

[47] Zhao, Z., Cui, T., Ding, S., Li, J., & Bellotti, A. G. (2024). Resampling Techniques Study on Class Imbalance Problem in Credit Risk Prediction. *Mathematics*, *12*(5), 701.

[48] Yang, Dongqi & Xiao, Binqing. (2024). Feature Enhanced Ensemble Modeling With Voting Optimization for Credit Risk Assessment. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3445499.

[49] Abidemi, Abiodun & Ajegbile, Mojeed & Ajegbile, Yusuff & Adedeji, Joy & Dada, Cecilia. (2023). A Deep Learning Prediction Model For Loan Default.

[50] Long, Zhi & Chen, Xiangzhou. (2023). Early warning research on enterprise carbon emission reduction credit risk based on deep learning model under unbalanced data. Frontiers in Energy Research. 11. 10.3389/fenrg.2023.1274425.

[51] Paudel, Sagun & Devkota, Bidur & Timilsina, Suresh. (2023). Multi-Class Credit Risk Analysis Using Deep Learning. Journal of Engineering and Sciences. 2. 82-87. 10.3126/jes2.v2i1.60399.

[52] Fan Yang, Yanan Qiao, Cheng Huang, Shan Wang, Xiao Wang, An Automatic Credit Scoring Strategy (ACSS) using memetic evolutionary algorithm and neural architecture search, Applied Soft Computing, Volume 113, Part A, 2021, 107871, ISSN 1568-4946.

[53] Zhuang, Yanyu & Wei, Hua. (2024). Design of a Personal Credit Risk Prediction Model and Legal Prevention of Financial Risks. IEEE Access. PP. 1-1. 10.1109/ACCESS.2024.3466192.

[54] Liang, Yancheng & Zhang, Jiajie & Li, Hui & Liu, Xiaochen & Hu, Yi & Wu, Yong & Zhang, Jinyao & Liu, Yongyan & Wu, Yi. (2023). DeRisk: An Effective Deep Learning Framework for Credit Risk Prediction over Real-World Financial Data. 10.48550/arXiv.2308.03704.

[55] Hung, Ming-Hung & Ku, Chao-Hsun & Chen, Kai-Ying. (2023). Application of Task-Aligned Model Based on Defect Detection. Automation. 4. 327-344. 10.3390/automation4040019.

[56] Olushola, Akinbusola & Mart, Joseph. (2024). Fraud Detection using Machine Learning. 10.14293/PR2199.000647.v1.

[57] A, Mrs. (2025). Online Payment Fraud Detection Using Machine Learning. INTERANTIONAL Journal Of Scientific Research In Engineering And Management. 09. 1-9. 10.55041/Ijsrem42092.

[58] John, Ada & Elly, Abill & Noah, Asher. (2025). Real-Time Fraud Detection Using Machine Learning Techniques.

[59] Chung, Jiwon & Lee, Kyungho. (2023). Credit Card Fraud Detection: An Improved Strategy for High Recall Using KNN, LDA, and Linear Regression. Sensors. 23. 7788. 10.3390/s23187788.

[60] Nobel, S.M.N., Swapno, S.M.M.R., Islam, M.R. *et al.* A machine learning approach for vocal fold segmentation and disorder classification based on ensemble method. *Sci Rep* **14**, 14435 (2024).

[61] Emi-Johnson, Oluwabukola & Nkrumah, Kwame & Folasole, Adetayo & Amusa, Tope. (2023). Optimizing Machine Learning for Imbalanced Classification: Applications in U.S. Healthcare, Finance, and Security. 10.5281