Applied Computer Science, vol. 20, no. 3, pp. 51–66
doi: 10.35784/acs-2024-28

Submitted: 2024-07-27 | Revised: 2024-08-27 | Accepted: 2024-09-02

Keywords: fraud detection, card-based financial systems, BiGru, BiLST, ensemble models, Machine Learning

Toufik GHRIB [0000-0001-7174-8962]*, *Yacine KHALDI* [0000-0002-8004-7698]*,
Purnendu Shekhar PANDEY [0000-0003-1276-5388]**,
Yusef Awad ABUSAL[0009-0000-3550-6384]***

ADVANCED FRAUD DETECTION IN CARD-BASED FINANCIAL SYSTEMS USING A BIDIRECTIONAL LSTM-GRU ENSEMBLE MODEL

Abstract

This article addresses the challenges of fraud in card-based financial systems and proposes effective detection and prevention strategies. By leveraging recent data analytics and real-time monitoring, the study aims to enhance transaction security and integrity. The authors review existing fraud detection methodologies, emerging trends, and the evolving tactics of fraudsters, emphasizing the importance of collaboration among financial institutions, regulatory agencies, and technology providers. Our proposed solution is an ensemble model combining Bidirectional Gated Recurrent Unit (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM) networks, designed to capture complex transactional patterns more effectively. Comparative analysis of six machine learning classifiers—AdaBoost, Naïve Bayes, Decision Tree, Logistic Regression, Random Forest, and Voting—demonstrates that our BiLSTM-BiGRU ensemble model outperforms traditional methods, achieving a fraud detection performance score of 89.22%. This highlights the advanced deep learning model's superior ability to enhance the robustness and reliability of fraud detection systems.

1. INTRODUCTION

Fraud detection in card-based finance is an essential component in the process of preserving the integrity and security of financial transactions. With the rapid expansion of digital banking and e-commerce, the volume and complexity of financial transactions have increased significantly, making fraud detection an increasingly challenging task (Bin Sulaiman et al., 2022, Teh et al., 2018). Fraudulent activities cause substantial financial losses and undermine consumer trust and the overall stability of financial systems. Classical machine learning models, such as Decision Trees and Support Vector Machines (SVM) (Klusowski & Tian, 2024, Valkenborg et al., 2023), have been extensively employed in fraud

* École Normale Supérieure de Ouargla, Mathemathics department, Algeria, ghrib.toufik@univ-ouargla.dz, yacine.khaldi@gmail.com
** Group of institutions, Department of CSE, KIET, India, purnendu.pandeyuma@gmail.com
*** Ufa Stat petroleum University, Ufa, yousef-abusal@mail.ru

detection because of their straightforwardness and efficiency (Poongodi et al., 2021, Cherkassky et al., 2004). However, these models often struggle with the imbalanced nature of fraud detection datasets (Wen et al., 2024), where the quantity of valid transactions much exceeds the number of fraudulent ones. To address this issue, the authors explore resampling techniques such as Random Under-sampling, Random Over-sampling, and Synthetic Minority Over-Sampling Technique (SMOTE) (Aghware et al., 2024) in order to achieve equilibrium in the datasets and enhance the performance of the model.

This study proposes a robust fraud detection framework that leverages an ensemble of advanced deep learning architectures, specifically Bidirectional Gated Recurrent Unit (BiGRU) and Bidirectional Long Short-Term Memory (BiLSTM) (Xu et al., 2023). These models are designed to capture complex patterns in transactional data more effectively than traditional approaches. By combining BiLSTM and BiGRU in an ensemble, the model benefits from the comprehensive contextual analysis of BiLSTM and the computational efficiency of BiGRU, achieving a robust balance between accuracy and resource efficiency (Stamate et al., 2024). The findings of the present experiment clearly show that the ensemble model significantly outperforms conventional machine learning models in terms of F1 score, Area Under the Curve (AUC), Recall, and Precision, particularly when combined with appropriate resampling techniques.

The structure of this article is as follows: The Literature Review section offers a comprehensive summary of current methodology and recent breakthroughs in fraud detection techniques. The Methods section details the machine learning models and deep learning architectures used in the study, including the ensemble model and resampling techniques applied to address class imbalance. The Results section displays the performance metrics of the models, illustrating the effectiveness of the proposed approach. Finally, the Discussion and Conclusion sections synthesize the findings, discuss the implications of the results, and propose future research directions to further enhance fraud detection in financial systems.

2. LITERATURE REVIEW

Sahin and Duman (Sahin et al., 2011) investigated how Support Vector Machines (SVM) and Decision Trees are employed for the purpose of identifying credit card fraud. They used decision trees, a machine learning approach that is simple to understand and apply. Because decision trees can manage large datasets and spot trends that can indicate fraudulent activity, they are highly beneficial in identifying instances of credit card fraud. To increase the model's predictive ability, the study also used support vector machines, a potent method for classification tasks. A dataset with credit card transactions classified as either fraudulent or legitimate was used in the study approach. To examine and categorize the transactions according to different criteria, the authors used SVM and decision trees. The outcomes showed that decision trees and SVM have the potential to be effective tools in order to detect credit card fraud due to their encouraging accuracy in identifying fraudulent activity. Patil et al. (2015) used a decision tree induction algorithm for the identification of credit card fraud. The authors suggest a method for using decision trees to improve the precision and effectiveness of fraud detection. The authors developed a model that could distinguish between authentic and fraudulent credit card transactions using a decision tree induction

approach. A dataset of past credit card transactions with labeled results is used to train the algorithm. The study highlights decision trees' interpretability, which makes them a sensible option for fraud detection systems where it is essential to comprehend the decision-making process. The study demonstrates how well the decision tree induction method performs in minimizing false positives and correctly recognizing fraudulent transactions. The results imply that decision trees, which strike a balance between interpretability and accuracy, can be a useful tool in credit card fraud detection systems.

In-depth research on credit card fraud detection methods by Sorournejad et al. (2016) offered a dual perspective by examining both the data-oriented and technique-oriented components. The survey, published in a respected journal, attempts to classify and evaluate current approaches based on the types of data used and the specific strategies applied. Techniques are categorized by the authors into groups such as hybrid approaches, machine learning algorithms, and rule-based methods. The survey provides a comprehensive overview of the topic and offers significant insights into the advantages and disadvantages of different techniques. This paper is an essential resource for researchers and practitioners who are interested in gaining an understanding of the wide range of strategies that may be used to detect fraudulent activity on credit cards.

The use of Generative Adversarial Networks (GANs) to improve classification efficacy in credit card fraud detection is investigated by Fiore et al. (2019). The researchers' innovative method enhanced the efficacy of the fraud detection model. The study focused on how well GANs generated fake examples of fraudulent transactions, which eventually resulted in a classification model that is more reliable and accurate. This creative application of GANs advances the capabilities of credit card fraud detection systems. An approach that is based on data mining for detecting credit card fraud in the e-commerce industry is presented by Carneiro et al. (2017). The research presented a comprehensive method for identifying patterns suggestive of fraud using techniques of data mining. In the context of online transactions, the authors stressed the importance of real-time detection. The suggested solution seeks to reduce false positives while increasing fraud detection accuracy through the use of data mining algorithms.

The authors in (Cui et al., 2021) presented a model focused on anomaly detection for online banking fraud, leveraging a technique known as multi-contextual behavior profiling. The proposed model, named ReMEMBeR, is designed to address several challenges in fraud detection, including limited historical behavior data, the heterogeneous nature of transaction data, and highly skewed class distributions between legitimate and fraudulent transactions. The model integrates information from similar users through a pseudo-recommender system approach and uses an embedding-based method to handle various attribute types uniformly. By using collaborative filtering and multi-contextual profiling, ReMEMBeR aims to distinguish fraudulent transactions from legitimate ones more effectively, enhancing both the accuracy and robustness of fraud detection systems. In Sudha et al. (2021), the authors developed a majority vote ensemble classifier aimed at improving the accuracy of credit card fraud detection. This approach integrates user behavior data, operational characteristics, and transactional details into a single feature set. The classification of user behaviors is performed using the Web Markov Skeleton Process (WMSP), operational features are analyzed with a Random Forest (RF) classifier, and transactional features are processed using SVM. The outputs from these individual classifiers are then combined through a majority voting ensemble (MVE) classifier, which leverages the strengths of each method

to enhance the overall detection accuracy. Halvaiee and Akbari (2014) presented a fraud detection approach using Artificial Immune Recognition System (AIRS), focusing on reducing training time through distributed computing. They used Hadoop's MapReduce framework to parallelize the training phase, specifically addressing the calculation of the affinity threshold and memory cell generation. The Map function handles distance calculations between records, and the Reduce function aggregates the results. This distributed implementation significantly accelerates the training process, making it suitable for handling large-scale transaction data efficiently. In Zhang et al. (2021), a novel feature engineering methodology was presented, named HOBA (Homogeneity-Oriented Behavior Analysis) for credit card fraud detection, incorporating deep learning architectures. They used a deep learning framework to model transaction behaviors more effectively by leveraging the HOBA framework, which focuses on generating feature variables that better capture the nuances of fraudulent behavior. The study's empirical results, based on data from a major commercial bank, demonstrated the superiority of this approach over traditional methods. The proposed system achieved better detection performance, including higher precision and recall rates, under various conditions. The authors in Zheng (2020) proposed a novel model called One-Class Adversarial Nets (OCAN) for fraud detection, designed to identify malicious users even when only benign user data is available. OCAN uses an LSTM-Autoencoder to encode the behavior of benign users into a latent space and a complementary GAN to generate samples that complement, rather than mimic, the benign data. The generator in the GAN creates these complementary samples, while the discriminator is trained to distinguish between actual benign data and these generated samples, effectively identifying anomalies. This approach eliminates the need for labeled malicious user data, making it highly adaptable and efficient for detecting fraud in dynamic, real-world scenarios.

3. PROPOSED METHOD

We proposed a deep learning ensemble model combining the best-performing state-of-the-art models: BiLSTM (GR & P, 2024, Wang, 2024) and BiGRU (Gorle & Panigrahi, 2023, Duarte Soares et al., 2022). The structure of the suggested ensemble model is illustrated in Fig.1. The model accepts two types of input: categorical and numerical data. The BiLSTM and BiGRU models offer distinct advantages for handling sequential data, which is essential for analyzing transaction patterns. BiLSTM networks excel due to their ability to process information in both forward and backward directions, enabling them to capture long-range dependencies and contextual details crucial for detecting subtle fraudulent activities. This bidirectional processing enhances the model's sensitivity to irregularities that may occur across various time frames. In contrast, BiGRU models present a more streamlined alternative with fewer parameters and reduced computational demands compared to LSTMs, yet they still effectively learn from sequential data. Combining BiLSTM and BiGRU in an ensemble model capitalizes on the strengths of both: the comprehensive context capture of BiLSTM and the efficient learning of BiGRU. This approach provides a balanced solution that achieves high accuracy while maintaining practical computational efficiency, ensuring the fraud detection system can effectively identify complex patterns without excessive resource use.

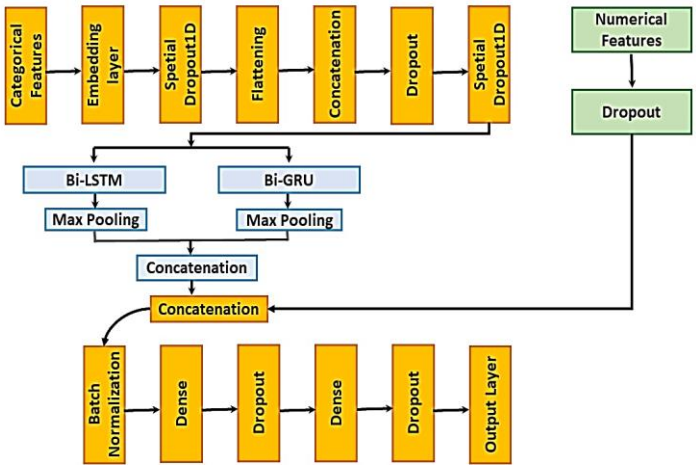

Fig.1. Proposed model's architecture

The suggested model combines the advantages of BiLSTM and BiGRU to effectively capture intricate patterns and interconnections in transactional data, offering a resilient method for detecting fraudulent activities.

1. Input Layers:

– Categorical Features: Categorical features are processed through an embedding layer to convert them into dense vectors. This is followed by a spatial Dropout1D layer to prevent overfitting and improve generalization.

– Numerical Features: Numerical features are directly passed through a dropout layer to mitigate overfitting.

2. Feature Processing:

– Embedding and Flattening: The embedded categorical features are flattened into a single dimension and concatenated with numerical features after applying dropout.

– BiLSTM and BiGRU Layers: The processed features are simultaneously fed into the BiLSTM and BiGRU layers. Both layers include max pooling to capture the most significant features from the sequences.

– Concatenation: The outputs from the BiLSTM and BiGRU layers are concatenated to combine the learned features from both architectures.

3. Dense and Dropout Layers:

– Batch Normalization: The concatenated output undergoes batch normalization to stabilize and accelerate the training process.

– Fully Connected Layers: Afterwards, there is a sequence of dense layers that are alternated with dropout layers. Dropout layers mitigate overfitting by randomly deactivating units throughout the training process.

– Output Layer: The last dense layer employs a sigmoid activation function to produce the output, indicating the probability of a transaction being fraudulent.

This ensemble approach leverages the bidirectional nature of BiLSTM and BiGRU to effectively capture the temporal dependencies and sequential patterns in the transaction data, which are crucial for accurately identifying fraudulent activities.

The proposed BiLSTM-BiGRU ensemble model offers several significant advantages for fraud detection in card-based financial systems. By combining the strengths of both BiLSTM and BiGRU architectures, the model robustly captures complex temporal dependencies and sequential patterns in transactional data, which are often missed by conventional machine learning models. By using dropout layers and batch normalization, the issue of overfitting is effectively addressed, resulting in improved generalization of the model to new, unseen data. This results in improved detection capabilities, as evidenced by superior performance metrics such as AUC, Precision, Recall, and F1 scores. Furthermore, the ensemble approach ensures a more comprehensive feature learning process, leveraging the bidirectional nature of both LSTM and GRU units to thoroughly analyze the data. These advantages collectively make the proposed model a powerful and reliable tool for enhancing the security and integrity of financial transactions.

4. EXPERIMENTAL ANALYSIS

4.1. Dataset and preprocessing

The dataset used in this experiment was the Credit Card Fraud Detection dataset, obtained from Kaggle (Machine Learning Group, 2024). The dataset comprises debit and credit card transactions conducted by European cardholders in September 2013. The dataset contains a total of 284,807 transactions, out of which 492 are classified as fraudulent operations. Although fraudulent transactions represent only 0.172% of the overall transactions, there is room for more balance in the statistics.

The dataset comprises thirty variables, which include:

1. Time: The time interval, measured in seconds, between the transaction and the initial transaction in the dataset.
2. V1 to V28: Principal components derived using a PCA transformation for the purpose of safeguarding sensitive data.
3. Amount: The monetary value of the transaction.
4. Class: The dependent variable, with a value of 1 indicating fraudulent transactions and a value of 0 indicating genuine transactions.

Data preprocessing is an essential and crucial stage in order to guarantee the quality and effectiveness of the machine learning model. The subsequent preprocessing procedures were implemented:

- Data Cleaning: The dataset was inspected for missing values. Since no missing values were in the dataset, no imputation was necessary.
- Scaling: Time and Amount features were not transformed using PCA and therefore required scaling. Both features were scaled using the standard scaler to normalize their distribution. This helps in improving the convergence of gradient-based algorithms.
- Data Imbalance Handling: Various strategies were considered to handle this imbalance. The approaches included undersampling the majority class, oversampling the minority class using SMOTE (Synthetic Minority Over-Sampling Technique), and experimenting with different class weights in the model.

To evaluate the performance of the machine learning models, the dataset was divided into training and testing sets, allowing for an assessment of their effectiveness. Eighty percent of

the data was allocated for training the model, while the remaining twenty percent was reserved for testing.

The splitting process was conducted in a stratified manner to ensure that the class distribution in both the training and testing sets appropriately reflected the imbalance that existed in the original dataset between fraudulent and non-fraudulent transactions.

The stratified split is crucial in this context due to the dataset's high imbalance, where only 0.172% of the transactions are fraudulent. Without stratification, there's a risk that the minority class (fraudulent transactions) might be underrepresented in either the training or testing set, which could lead to biased model performance metrics.

4.2. Performance metrics

In order to assess the efficacy of the machine learning models in identifying credit card fraud, multiple performance indicators were used. Due to the significant imbalance in the dataset, it is essential to use measurements that surpass basic accuracy. The subsequent metrics were used: The concepts of Area Under the Curve (AUC), Precision, Recall, and F1 score are important in evaluating the performance of a model.

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a metric used to evaluate the performance of classification models across different threshold values. The AUC quantifies the level of distinctiveness, reflecting the model's ability to differentiate between classes. A receiver operating characteristic (ROC) curve with an area under the curve (AUC) value of 1.0 signifies a flawless model, whereas an AUC of 0.5 signifies a model that lacks the ability to discriminate.

The AUC-ROC is especially valuable when dealing with imbalanced datasets as it offers a comprehensive assessment of performance across all categorization levels. Precision is the quotient obtained by dividing the number of accurately anticipated positive observations by the total number of predicted positive observations. The query aims to determine the number of fraudulent transactions among all the flagged transactions.

$$Precision = TP/(TP + FP) \tag{1}$$

TP represents the count of correctly identified positive instances, while FP represents the count of incorrectly identified positive instances. High precision is crucial in fraud detection as it ensures a low rate of false positives, minimizing the number of normal transactions that are mistakenly identified as fraudulent.

Recall, often referred to as Sensitivity or True Positive Rate, is the proportion of accurately predicted positive observations to the total number of observations in the actual class. The question it addresses is: what is the number of accurately identified fraudulent transactions out of the total number of fraudulent transactions?

$$Recall = TP/(TP + FN) \tag{2}$$

FN represents the quantity of false negatives. Maximizing recall is essential in fraud detection to ensure the identification of the majority of fraudulent transactions, even if it results in the inclusion of some false positives.

The F1 score is calculated as the harmonic mean of the Precision and Recall metrics. The metric offers a unified measure that effectively manages the trade-off between Precision and

Recall, which is particularly crucial in situations where there is an imbalance between the two.

$$F1Score = 2 * (Precision * Recall)/(Precision + Recall) \quad (3)$$

The F1 score is especially valuable when working with imbalanced datasets since it provides a more accurate measure of the erroneously categorized cases compared to the accuracy metric.

4.3. RESULT AND DISCUSSION

The primary objective of the initial experiment was to evaluate the effectiveness of several machine learning algorithms in handling the problem of class imbalance in the credit card fraud detection dataset. This was achieved using an oversampled dataset. The use of the oversampling strategy guarantees sufficient representation of the minority class (fraudulent transactions), hence enhancing the model's capacity to identify fraud.

Tab.1 presents the outcomes of applying several machine learning models, including Naïve Bayes, Voting Classifier, Random Forest, Logistic Regression, AdaBoost, and Decision Tree, to the oversampled dataset.

Tab. 1. The outcomes of applying Machine Learning models to an oversampled dataset

Model	AUC	Precision	Recall	F1_score
Naïve base	61.25	80.11	88.12	62.25
Voting	69.33	75.36	85.36	71.33
Random Forest	68.15	60.25	71.22	85.25
logistic regression	71.22	77.15	69.25	75.36
Ada boosting	79.25	83.21	68.15	88.15
Decision Tree	80.21	89.11	70.21	79.23

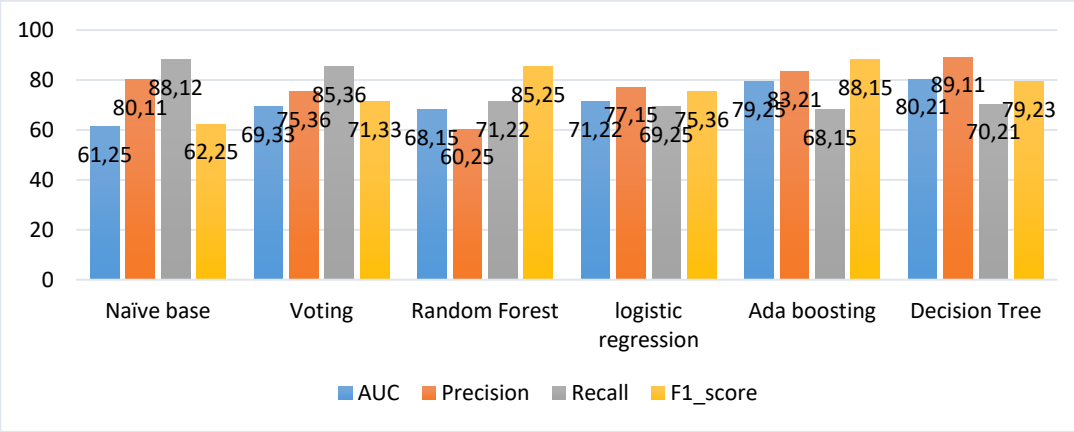

Fig.2. The outcomes of applying Machine Learning models to an oversampled dataset

The Naive Bayes model has a moderate AUC of 61.25, indicating that it discriminates between classes rather well. It had a high Precision of 80.11%, indicating a low false positive rate, and a respectable Recall of 88.12%, demonstrating its ability to identify a significant

proportion of real positive occurrences. The resulting F1 score of 62.25 demonstrates a good balance of Precision and Recall.

The Voting model, which combined multiple classifiers, achieved a better AUC of 69.33. It had a Precision of 75.36%, indicating reasonable accuracy in positive predictions, and a Recall of 85.36%, demonstrating its ability to capture a considerable portion of positive instances. The model's balanced performance was further emphasized by its F1 score of 71.33.

The Random Forest model achieved a competitive AUC of 68.15. While its Precision was 60.25%, showing a moderate degree of accurate positive predictions, its Recall of 71.22% indicated that it correctly predicted a significant proportion of real positive instances. Surprisingly, the high F1 score of 85.25 indicated a solid combination of Precision and Recall.

The AUC of Logistic Regression was 71.22, demonstrating its ability to differentiate across classes. It had remarkable accuracy in positive predictions with a Precision of 77.15%, but a Recall of 69.25% indicated that it may have missed some positive examples. The F1 score of 75.36 demonstrated a healthy balance of Precision and Recall.

Ada-Boosting outperformed other models, with an AUC of 79.25. It had a high Precision of 83.21%, showing a low false positive rate, but a poor Recall of 68.15%, indicating a trade-off with false negatives. The model's overall effectiveness in establishing a balance between Precision and Recall was underlined by its remarkable F1 score of 88.15.

Finally, the Decision Tree model outperformed with an AUC of 80.21. It had an excellent Precision of 89.11%, suggesting great accuracy in positive predictions, and a Recall of 70.21%, capturing a significant amount of positive events. The F1 score of 79.23 demonstrated a strong balance of Precision and Recall. In conclusion, the Decision Tree model performed well in this examination, demonstrating its ability to handle the given dataset.

The second experiment aimed to assess the efficacy of different machine learning models using an under-sampled dataset to mitigate the issue of class imbalance. Tab.2 presents a concise overview of the results obtained from this experiment:

- Naïve Bayes: Achieved an AUC of 82.12, but with lower Precision (39.25) and Recall (51.21), resulting in a low F1 score of 31.25. This indicates poor performance in identifying fraudulent transactions accurately.
- Voting Classifier: Showed an AUC of 78.33, with moderate Precision (71.22) and low Recall (36.22), leading to an F1 score of 44.15. This model struggled with Recall, indicating many missed frauds.
- Random Forest: Reported an AUC of 75.15, with high Precision (85.36) but lower Recall (61.25) and an F1 score of 36.25. The high precision indicates fewer false positives, but the recall suggests many fraudulent transactions were not detected.
- Logistic Regression: Achieved an AUC of 82.33, with balanced Precision (71.22) and Recall (71.25), resulting in an F1 score of 48.66. This model provided a balanced performance but still lacked in F1 score.
- AdaBoost: Recorded an AUC of 80.18, with good Precision (80.65) and high Recall (85.11), resulting in an F1 score of 51.23. It performed well in identifying fraudulent transactions with a balanced approach.

– Decision Tree: Achieved an AUC of 79.32, with balanced Precision (79.23) and Recall (79.21), resulting in the highest F1 score of 61.23 among all models tested on the undersampled dataset.

The results indicate that while some models, like Decision Tree and AdaBoost, managed to maintain a good balance between Precision and Recall, others like Naïve Bayes and Voting Classifier struggled with low F1 scores, primarily due to lower Recall. This experiment highlights the challenge of achieving high performance on an undersampled dataset, where maintaining a balance between detecting fraudulent transactions and minimizing false positives is crucial.

Tab.2. The outcomes of applying Machine Learning models to an undersampled dataset

Model	AUC	Precision	Recall	F1 score
Naïve base	82.12	39.25	51.21	31.25
Voting	78.33	71.22	36.22	44.15
Random Forest	75.15	85.36	61.25	36.25
logistic regression	82.33	71.22	71.25	48.66
Ada boosting	80.18	80.65	85.11	51.23
Decision Tree	79.32	79.23	79.21	61.23

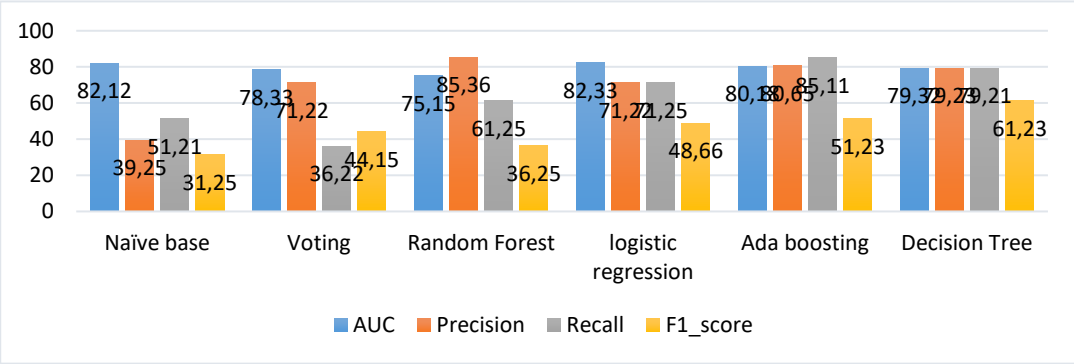

Fig.3. The outcomes of applying Machine Learning models to an unsampled dataset

The third experiment entailed the use of machine learning models on the dataset, which was balanced using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a widely used technique for addressing imbalanced datasets by creating artificial samples for the underrepresented class. Table 3 presents a concise overview of the performance measures for different models.

Tab.3. The outcomes of Machine Learning models with SMOTE

Model	AUC	Precision	Recall	F1_score
Naïve base	68.21	11.23	11.2	18.2
Voting	56.21	25.36	16.2	35.1
Random Forest	71.22	34.12	18.3	41.2
logistic regression	66.25	54.22	20.3	22.6
Ada boosting	70.25	62.14	29.5	35.5
Decision Tree	69.12	60.36	31.2	41.6

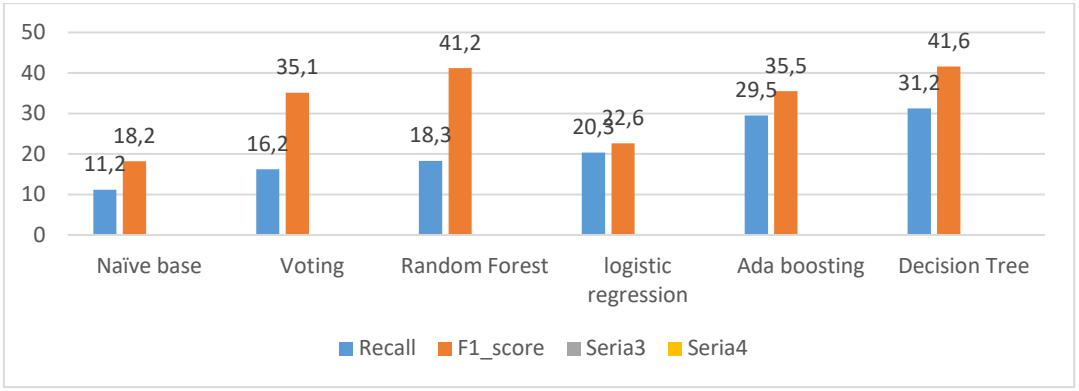

Fig.4. The outcomes of applying Machine Learning models to smote

Naïve Bayes achieved an AUC of 68.21, with low Precision (11.23) and Recall (11.2), resulting in an F1 score of 18.2. This indicates poor performance with a high number of false positives. The Voting Classifier showed an AUC of 56.21, with moderate Precision (25.36) and low Recall (16.2), leading to an F1 score of 35.1. This model struggled with both precision and recall. Random Forest reported an AUC of 71.22, with better Precision (34.12) and Recall (18.3), resulting in an F1 score of 41.2. This model improved in identifying fraudulent transactions compared to previous models. Logistic Regression achieved an AUC of 66.25, with higher Precision (54.22) and Recall (20.3), resulting in an F1 score of 22.6. AdaBoost recorded an AUC of 70.25, with good Precision (62.14) and Recall (29.5), resulting in an F1 score of 35.5. It performed well in identifying fraudulent transactions with a balanced approach. Decision Tree achieved an AUC of 69.12, with balanced Precision (60.36) and Recall (31.2), resulting in the highest F1 score of 41.6 among all models tested with SMOTE.

Table 4 displays the results of assessing the effectiveness of two deep learning models, BiLSTM and BiGRU, on an unbalanced dataset. The evaluation was conducted using three distinct sampling techniques: random undersampling, random oversampling, and SMOTE.

1. BiLSTM:

On the imbalanced dataset, BiLSTM achieved an AUC of 88.00% and an accuracy of 87.4%. With random undersampling, BiLSTM improved slightly, achieving an AUC of 88.15% and an accuracy of 89.6%. With random oversampling, BiLSTM achieved an AUC of 89.39% and an accuracy of 90.2%, showing better performance in identifying fraudulent transactions. Using SMOTE, BiLSTM achieved a lower AUC of 82.12% but the highest accuracy of 98.5%, indicating a significant improvement in overall classification accuracy.

2. BiGRU:

When working with an imbalanced dataset, BiGRU fared marginally better than BiLSTM, with an AUC of 89.07% and an accuracy of 90.8%. By employing random undersampling, the BiGRU model achieved an AUC of 88.15%, which was equivalent to the AUC of the BiLSTM model. Additionally, the BiGRU model demonstrated an accuracy of 89.6%. By employing random oversampling, the BiGRU model achieved comparable results to the BiLSTM model, exhibiting an AUC (Area Under the Curve) of 89.39% and an accuracy of 90.2%. When SMOTE was used, BiGRU had a decrease in AUC to 82.12%, but achieved the maximum accuracy of 98.5%, which was the same as BiLSTM.

Tab.4. Deep Learning model results using the three sampling techniques

	Imbalanced dataset		Random under-sampling		Random over-sampling		SMOTE	
	AUC	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**	**ACC**
BiLSTM	88.00%	87.4%	88.15%	89.6%	89.39%	90.2%	82.12%	98.5%
BiGRU	89.07%	90.8%	91.01%	91.52%	91.24%	90.15%	84.36%	90.02%

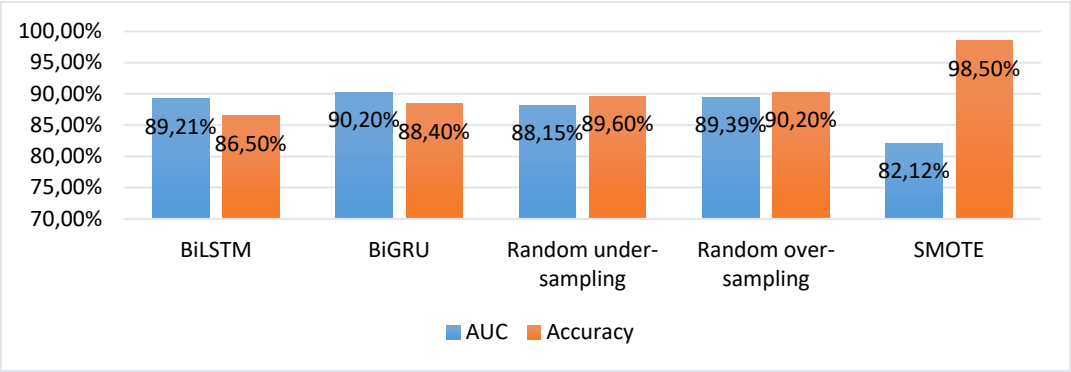

Fig. 5. Deep Learning model results using the three sampling techniques

Table 5 presents the performance metrics of the proposed BiLSTM-BiGRU-based ensemble model when applied to datasets balanced using three different sampling techniques:

- Random Undersampling resulted in the ensemble model achieving an AUC of 91.23%, Precision of 89.21%, Recall of 82.29%, and an F1 score of 87.51%. The results demonstrate that the model has outstanding performance in differentiating between fraudulent and non-fraudulent transactions, with a robust balance between precision and recall.
- The model achieved an AUC (Area Under the Curve) of 89.22%, a Precision of 65.21%, a Recall of 88.14%, and an F1 score of 88.25% using Random Oversampling. Although the AUC is slightly lower compared to random undersampling, the model demonstrates high recall and F1 score, indicating its effectiveness in spotting fraudulent transactions. However, this comes at the cost of some precision.
- The SMOTE model produced an Area Under the Curve (AUC) of 85.21%, a Precision of 69.35%, a Recall of 69.26%, and an F1 score of 75.36%. While the accuracy is notably superior in comparison to the other sampling approaches, the lower AUC and F1 score suggest that the model may not exhibit strong generalization when applied to unfamiliar data.

These results highlight that the ensemble model of BiLSTM and BiGRU performs best with random undersampling, achieving the highest AUC and a good balance between precision and recall. Random oversampling also provides strong performance, particularly in terms of recall and F1 score, making it a suitable technique for scenarios where identifying as many fraudulent transactions as possible is critical. However, SMOTE, while improving overall classification accuracy, results in lower AUC and F1 scores, suggesting a potential overfitting issue.

Tab. 5. Proposed model's results

Technique	AUC	Precision	Recall	F1_score
Random under-sampling	91.23%	89.21%	82.29%	87.51%
Random over-sampling	89.22%	65.21%	88.14%	88.25%
SMOTE	85.21%	69.35%	69.26%	75.36%

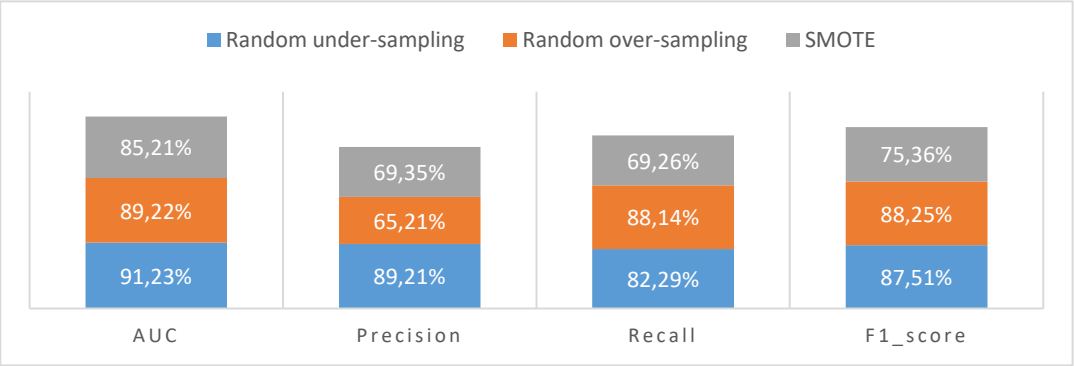

Fig. 6. The outcomes of our model

As shown in Tab.6, the authors evaluated the performance of their proposed BiLSTM-BiGRU ensemble model in the identification of fraudulent activity by comparing it to methods that are considered to be state-of-the-art. When it comes to dealing with the intricacies of fraud detection in card-based financial systems, the suggested BiLSTM-BiGRU ensemble model reveals significant gains over methods that are considered to be as advanced as they are currently.

Tab. 6. Performance comparison of BiLSTM-BiGRU ensemble model with state-of-the-art methods

Article	Method	Dataset	AUC	Precision	Recall	F1 score
(Cui et al., 2021)	ReMEMBeR	Real-world online banking transaction dataset	-	86.42%	81.69%	82.48%
(Sudha et al., 2021)	Majority Vote Ensemble Classifier	Bestpay digital payment platform dataset	-	86%	94%	90%
(Halvaiee & Akbari, 2014)	AIS-based Fraud Detection Model	Transaction data from a Brazilian bank	-	56%	51%	-
(Zhang et al., 2021)	Feature Engineering Methodology	Real-life dataset from a bank in China	-	62.60%	75%	57.7%
(Zheng, 2020)	Generative Adversarial Networks	Synthetic data	-	81.98%	55.69%	65.37%
This article	BiSTM-BiGRU ensemble model	Credit Card Fraud Detection	91.23%	89.21%	82.29%	87.51%

5. CONCLUSION

Maintaining the integrity and security of financial transactions is of utmost importance in the detection and prevention of fraud in card-based financial systems. Given the intricate and interrelated nature of today's financial system, it is imperative to employ strong fraud detection methods to effectively combat fraudulent actions. This study involved an investigation of different methods and techniques, such as machine learning models, resampling techniques, and deep learning architectures.

It is crucial to achieve a harmonious equilibrium among accuracy, precision, recall, and the general efficiency of the system. Random Undersampling and Random Oversampling are effective techniques for addressing class imbalances, whereas deep learning architectures such as BiLSTM and BiGRU have advanced capacities to capture intricate patterns in transactional data. Incorporating resampling approaches, such as SMOTE, is beneficial for constructing a complete fraud detection framework as it delivers significant insights.

The BiLSTM-BiGRU ensemble model suggested by the authors outperformed typical machine learning classifiers, demonstrating higher metrics across key performance indices. The ensemble model's capacity to properly balance precision and recall underscores its resilience and potential for practical financial applications.

Financial institutions can enhance the security of their card-based systems by implementing these measures, thereby protecting clients and ensuring the integrity of the financial ecosystem. In order to retain the confidence and security of card-based financial transactions, it is crucial to stay updated with technical breakthroughs and continuously improve fraud detection methods as the financial landscape evolves. Future research and development in the field should focus on implementing real-time fraud detection systems, improving the ability to explain and understand the models used, integrating various sources of data, creating adaptive learning systems, fostering collaboration and information sharing among financial institutions, and addressing regulatory and ethical concerns to ensure privacy and fairness in fraud detection. By embracing these viewpoints, the domain of fraud detection can make substantial progress, offering more robust and trustworthy financial systems that can effectively combat the always evolving realm of fraudulent activity.

Acknowledgment

The authors express their gratitude to the director of ENS Ouargla, Algeria for facilitating access to the structures and equipment. This publication is a component of the PRFU project, specifically assigned the number D04N01EN300120220001. The Algerian Ministry of Higher Education and Scientific Research

REFERENCES

Aghware, F. O., Ojugo, A. A., Adigwe, W., Odiakaose, C. C., Ojei, E. O., Ashioba, N. C., Okpor, M. D., & Geteloma, V. O. (2024). Enhancing the random forest model via synthetic minority oversampling technique for credit-card fraud detection. *Journal of Computing Theories and Applications*, *1*(4), 407-420. https://doi.org/10.62411/jcta.10323

Bin Sulaiman, R., Schetinin, V., & Sant, P. (2022). Review of machine learning approach on credit card fraud detection. *Human-Centric Intelligent Systems*, *2*, 55-68. https://doi.org/10.1007/s44230-022-00004-0

Carneiro, N., Figueira, G., & Costa, M. (2017). A data mining based system for credit-card fraud detection in e-tail. *Decision Support Systems*, *95*, 91-101. https://doi.org/10.1016/j.dss.2017.01.002

Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural networks*, *17*(1), 113-126. https://doi.org/10.1016/S0893-6080(03)00169-2

Cui, J., Yan, C., & Wang, C. (2021). ReMEMBeR: Ranking metric embedding-based multicontextual behavior profiling for online banking fraud detection. *IEEE Transactions on Computational Social Systems*, *8*(3), 643-654. https://doi.org/10.1109/TCSS.2021.3052950

Duarte Soares, L., de Souza Queiroz, A., López, G. P., Carreño-Franco, E. M., López-Lezama, J. M., & Muñoz-Galeano, N. (2022). BiGRU-CNN neural network applied to electric energy theft detection. *Electronics*, *11*(5), 693. https://doi.org/10.3390/electronics11050693

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, *479*, 448-455. https://doi.org/10.1016/j.ins.2017.12.030

Gorle, V. L. N., & Panigrahi, S. (2023). A semi-supervised Anti-Fraud model based on integrated XGBoost and BiGRU with self-attention network: an application to internet loan fraud detection. *Multimedia Tools and Applications*, *83*, 56939–56964. https://doi.org/10.1007/s11042-023-17681-z

GR, J., & P, A. I. (2024). Attention layer integrated BiLSTM for financial fraud prediction. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-024-18764-1

Halvaiee, N. S., & Akbari, M. K. (2014). A novel model for credit card fraud detection using Artificial Immune Systems. *Applied Soft Computing*, *24*, 40-49. https://doi.org/10.1016/j.asoc.2014.06.042

Klusowski, J. M., & Tian, P. M. (2024). Large scale prediction with decision trees. *Journal of the American Statistical Association*, *119*(545), 525-537. https://doi.org/10.1080/01621459.2022.2126782

Machine Learning Group. (2024). *Credit card fraud detection*. Retrieved May 5, 2024 from https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Patil, S., Somavanshi, H., Gaikwad, J. B., Deshmane, A., & Badgujar, R. (2015). Credit card fraud detection using decision tree induction algorithm. *International Journal of Computer Science and Mobile Computing*, *4*(4), 92-95.

Poongodi, K., & Kumar, D. (2021). Support vector machine with information gain based classification for credit card fraud detection system. *The International Arab Journal of Information Technology*, *18*(2), 199-207. https://doi.org/10.34028/iajit/18/2/8

Sahin, Y., & Duman, E. (2011). Detecting credit card fraud by decision trees and support vector machines. *International MultiConference of Engineers and Computer Scientists 2011 (IMECS 2011)* (pp. 1-6).

Sorournejad, S., Zojaji, Z., Atani, R. E., & Monadjemi, A. H. (2016) A survey of credit card fraud detection techniques: Data and technique oriented perspective. *ArXiv, abs/1611.06439*. https://doi.org/10.48550/arXiv.1611.06439

Stamate, D., Davuloori, P., Logofatu, D., Mercure, E., Addyman, C., & Tomlinson, M. (2024). Ensembles of vidirectional LSTM and GRU neural nets for predicting mother-infant synchrony in videos. In L. Iliadis, I. Maglogiannis, A. Papaleonidas, E. Pimenidis, & C. Jayne (Eds.), *Engineering Applications of Neural Networks* (Vol. 2141, pp. 329–342). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-62495-7_25

Sudha, C., & Akila, D. (2021). WITHDRAWN: Majority vote ensemble classifier for accurate detection of credit card frauds. *Materials Today: Proceedings*. https://doi.org/10.1016/j.matpr.2021.01.616

Teh, B., Islam, M. B., Kumar, N., Islam, M. K., & Eaganathan, U. (2018). Statistical and spending behavior based fraud detection of card-based payment system. *2018 International Conference on Electrical Engineering and Informatics (ICELTICs)* (pp. 78-83). IEEE. https://doi.org/DOI:10.1109/ICELTICS.2018.8548878

Valkenborg, D., Rousseau, A. J., Geubbelmans, M., & Burzykowski, T. (2023). Support vector machines. *American Journal of Orthodontics and Dentofacial Orthopedics*, *164*(5), 754-757. https://doi.org/10.1016/j.ajodo.2023.08.003

Wang, S. (2024). Intelligent BiLSTM-Attention-IBPNN method for anomaly detection in financial auditing. *IEEE Access*, *12*, 90005-90015. https://doi.org/10.1109/ACCESS.2024.3420243

Wen, J., Tang, X., & Lu, J. (2024). An imbalanced learning method based on graph tran-smote for fraud detection. *Scientific Reports*, *14*, 16560. https://doi.org/10.1038/s41598-024-67550-4

Xu, L., Xu, W., Cui, Q., Li, M., Luo, B., & Tang, Y. (2023). Deep heuristic evolutionary regression model based on the fusion of BiGRU and BiLSTM. *Cognitive Computation*, *15*, 1672-1686. https://doi.org/10.1007/s12559-023-10135-6

Zhang, X., Han, Y., Xu, W., & Wang, Q. (2021). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. *Information Sciences*, *557*, 302-316. https://doi.org/10.1016/j.ins.2019.05.023

Zheng, P., (2020). Dynamic Fraud Detection via Sequential Modeling. *Graduate Theses and Dissertations*. Retrieved from https://scholarworks.uark.edu/etd/3633