Addressing Class Imbalance in Credit Card Fraud Detection: A Hybrid Deep Learning Approach

Muhammad Jabir Khan[1], Syed Irfan Ullah[1]

[1]Department of Computing and Technology, Abasyn University Peshawar, Pakistan.

* Correspondence: Muhammad Jabir Khan (muhammadjabirkhan@gmail.com).

Citation | Khan. M. J, Ullah. S. I, "Addressing Class Imbalance in Credit Card Fraud Detection: A Hybrid Deep Learning Approach", IJIST, Vol. 7 Issue. 1 pp 603-622, Mar 2025

Received | March 2, 2025 **Revised |** March 22, 2025 **Accepted |** March 24, 2025 **Published |** March 29, 2025.

The rise of credit card fraud is a global concern, demanding reliable detection methods that can overcome challenges with imbalanced datasets and limited exploration of hybrid modeling approaches. This study introduces a hybrid deep learning architecture combining Gated Recurrent Unit (GRU) and Long Short-Term Memory (LSTM) layers alongside SMOTE-TOMEK preprocessing to address imbalanced data issues in credit card fraud detection. The research analyzes a substantial dataset containing both legitimate and fraudulent transactions, evaluating the performance of GRU, LSTM, and the novel Hybrid model through comprehensive data exploration, preprocessing, and feature selection. Performance evaluation uses metrics including accuracy, precision, recall, F1 Score, AUROC, and AUPRC. The experimental results demonstrate the effectiveness of deep learning architectures, with AUROC values of 0.974551 for LSTM, 0.958174 for GRU, and 0.976205 for the Hybrid model. The Hybrid model showed particularly promising results with a precision of 0.9121 and AUPRC of 0.886068, outperforming the individual models. These findings indicate that combining complementary deep learning architectures enhances fraud detection by leveraging their respective strengths in capturing both long-term dependencies and transaction patterns. These insights offer valuable guidance to financial institutions in implementing effective fraud detection systems while emphasizing the importance of continuous improvement of deep learning algorithms to address evolving cyber threats.

Keywords: Credit card fraud detection; Deep Learning; GRU; LSTM; Smote-Tomek.

Introduction:

Credit card fraud has emerged as a critical concern, imperiling the financial security of a vast array of personal and corporate stakeholders worldwide. In 2021, the United States alone witnessed staggering losses of around $11.91 billion due to this issue. However, according to the Nilson Report published in December 2022, there is a slight decline in credit card fraud trends. The report projects that global losses from card fraud for retailers, acquirers, and issuers will amount to approximately $397.40 billion over the next decade, causing a notable reduction from the previous estimates of $408.50 billion [1].

Recently, there has been a growing interest in utilizing machine learning methods for identifying and preventing credit card fraud [2]. As credit card usage continues to grow, the need for reliable fraud detection methods becomes increasingly important. Machine learning techniques have emerged as a promising solution for addressing credit card fraud [3]. These techniques are capable of examining large volumes of transaction data to detect potentially suspicious patterns. Several approaches have been employed to effectively mitigate these frauds, with the prevalence of machine learning strategies due to their effectiveness in the quick identification of fraudulent transactions [4]. Decision Trees, Regression, Random Forests, Hidden Markov Models, and Genetic Algorithms lie among algorithms that have demonstrated high efficacy in detecting fraudulent transactions [5]. With the help of these techniques, financial institutions can have the ability to substantially reduce the probability of fraudulent credit card activities and provide a safe and secure transaction environment for their customers.

The current study focuses on employing the Smote-Tomek technique, notable for its efficacy in managing imbalanced datasets, to enhance fraud detection accuracy. Utilizing a comprehensive dataset comprising legitimate and fraudulent transactions, the primary goal is to develop hybrid deep learning models. These models aim to leverage diverse deep learning algorithms and integrate the Smote-Tomek technique for balancing the dataset, thereby enhancing model performance.

Research suggests that Smote-Tomek outperforms other techniques within the SMOTE family [6]. This underscores its suitability for applications requiring robust handling of class imbalance, particularly in fraud detection.

Objectives:

The main aim of this study is to enhance current approaches for detecting credit card fraud by utilizing an integrated approach encompassing sophisticated deep learning algorithms. This research seeks to evaluate the effectiveness of modern deep learning techniques combined with the Smote-Tomek method in detecting fraudulent activities. The researchers aim to provide crucial and in-depth insights that will be invaluable to financial institutions, guiding them in selecting the most accurate and efficient algorithms for fraud detection. The key contribution of the present research is as follows:

- A feature selection pipeline was developed, utilizing Random Forest importance scores and Logistic Regression, to determine the most identifying variables for identifying fraud with credit cards.
- The class imbalance challenge was addressed through a combined approach of Random sampling and SMOTE-Tomek, creating a more balanced dataset for model training.
- Three distinct recurrent neural network architectures were implemented and compared: an LSTM network, a GRU network, and a novel hybrid LSTM-GRU model.
- Model effectiveness was rigorously analyzed using an extensive set of parameters, including precision, recall, F1 score, and area under the ROC and Precision-Recall curves.
- A sophisticated hybrid model was designed, incorporating bidirectional LSTM and GRU layers with dropout regularization, demonstrating enhanced fraud detection capabilities compared to single-architecture models.

- Performance visualization techniques, including ROC and Precision-Recall curves, were employed to conduct a nuanced analysis of model behavior across various classification thresholds.

Literature Review:

This literature review examines existing research that explores the application of deep learning and machine learning methods in detecting credit card fraud. E.F. Malik et al. [7] have recently proposed new hybrid machine-learning approaches that aim to identify instances of credit card fraud. These approaches utilized advanced machine learning techniques in two phases. During the first phase, multiple machine-learning algorithms were utilized to detect cases of fraudulent activity involving credit cards. In the second phase, a combination of approaches was developed, using the top-performing algorithm from the previous phase as a foundation. The researchers investigated more than five hybrid machine-learning models using a real-world dataset. The findings indicated that the combined LightGBM and Adaboost approach exhibited the most superior performance with ROC 0.82. This model was able to detect unauthorized credit card transactions with remarkable accuracy, improving recall and precision scores of the previously used models. The results of this research could aid in the creation of improved approaches for identifying credit card fraud and stopping monetary loss.

S. K. Hashemi et al. proposed a novel method to detect fraudulent activities in banking data. To tackle the problem of unequal class distribution, they assessed how class weight-adjusting hyperparameters could be used to equalize the weight between genuine and fraudulent transactions. They also explored the application of Bayesian optimization to fine-tune the hyper-parameters, considering real-world factors like imbalanced data. The proposed approach has shown promising results in detecting fraudulent activities with an AUC of 0.952 and the proposed LightGBM with an AUC of 0.947 and can potentially be used in real-world banking systems to combat financial crimes [8].

In a study recently conducted by A. N. Ahmed et al., various machine learning algorithms were evaluated for their ability to detect fraud in credit cards. The tested algorithms included Extreme Gradient Boosting (XGBoost), and K-Nearest Neighbors (KNN), among others. The study revealed that ensemble-based methods such as XGBoost and Random Forest outperformed other algorithms in the detection of credit card fraud. This is significant as the proliferation of credit card-related fraudulent activities presents a persistent challenge, precipitating pecuniary losses that impact both private individuals and corporate entities. A. N. Ahmed et al. investigated a range of machine learning approaches like KNN, XGBoost, Naive Bayes, SVM, Logistic Regression, and Random Forest for the identification of fraud in credit cards. They found ensemble methods like XGBoost (AUC: 97.89%) and Random Forest (AUC: 97.89%) to be the most effective [9].

The increasing volume of sensitive data stored online has elevated cybersecurity to a critical concern in recent times. E. Jayanthi et al. conducted an investigation that utilized novel machine learning methods to enhance cybersecurity, to address these challenges [10]. The performance of these methods was compared with others using metrics comparison, and CCRF (Cluster and Classifier-based Random Forest) and CCLR (Cluster and Classifier-based Logistic Regression) methods were found to outperform other methods with accuracy rates of 99.95% and 99.97%. In today's digital economy, the use of credit cards is rapidly increasing, but unfortunately, so is credit card fraud. Traditional machine learning methods for fraud detection often fail due to evolving user behaviors and class imbalance. To tackle this problem, Mienye and Sun proposed recent deep-learning strategies specifically designed to overcome these limitations [11]. These strategies seek to enhance fraud detection by integrating deep learning techniques, which are more adept at managing evolving user behavior and addressing the class imbalance issues that traditional methods often struggle to overcome.

Vesta dataset and deep learning system is utilized to identify fraudulent credit card transactions. Their method achieved an impressive 99.1% ROC curve score, exhibiting its efficacy in detecting fraudulent activities [12]. Additionally, S. Kumar et al. proposed an intelligent system employing machine learning to uncover credit card fraud, which utilized an SVM classification method. Their method showed promising results in terms of accuracy, outperforming other existing techniques for detecting fraudulent transactions [13]. K. Jegadeesan et al. proposed an ensemble ML strategy for the detection of credit card fraud. They used SMOTE, Recursive Feature Elimination, and Ensemble Classifiers to tackle imbalanced data and identify optimal prediction features [14].

The field of credit card fraud analysis is being progressively transformed by incorporating cutting-edge deep learning models, notably, CNNs possess the capacity to automatically recognize and gather key characteristics from unprocessed data. Moreover, CNNs are adept at handling considerable amounts of data and can effectively adapt to new data. The robustness of deep learning frameworks is inherently linked to the comprehensiveness and veracity of the training data, along with the hyperparameters used during model training. To improve the efficiency of deep learning models in detecting fraudulent transactions, researchers have proposed various techniques for optimizing hyperparameters and pre-processing imbalanced data [15]. To effectively mitigate the effects of skewed class representation in credit card fraud analysis, Strelcenia and Prakoonwit (2023) introduced a novel GAN-based data augmentation method. Their approach, known as the K-CGAN method, outperformed traditional techniques like SMOTE and ADASYN by generating high-quality test datasets, which improved parameters like recall, F1-score, accuracy, and precision [16].

Similarly, Alabrah (2023) developed an improved credit card fraud detection system that incorporates outlier normalization utilizing IQR methodology for outlier detection and implementing SMOTEN-based oversampling. This approach resulted in a significant enhancement in the model's AUC score, achieving an AUC of 1. 00 [17]. Additionally, Mahajan and Baghel (2023) explored the use of logistic regression combined with under and oversampling strategies for addressing the challenge of class imbalance, achieving a 94% detection accuracy rate for fraudulent transactions [18]. Feldman et al. further contributed by employing the Tomek links, which significantly improved the reliability of fraud detection models [19].

The rise of credit card fraud has encouraged researchers to examine different machine learning (ML) techniques to improve detection. Hybrid models that combine approaches like Adaboost and LGBM have achieved high accuracy. Methods such as class weight-tuning and Bayesian optimization have been used to tackle imbalanced data. Deep learning strategies have addressed the limitations of traditional methods, and some have achieved high ROC scores. However, techniques like SMOTE and ensemble classifiers, useful for feature selection, may introduce noise or bias, which can compromise model performance. Deep learning models, particularly Convolutional Neural Networks (CNNs), hold significant potential but are highly dependent on data quality and effective hyperparameter optimization. This review underscores the significant potential of advanced ML techniques and artificial neural networks for mitigating credit card fraudulent activities. It also emphasizes the necessity for improved methods to handle imbalanced datasets without introducing bias, which is the focus of the proposed research.

Material and Methods:

Data Acquisition and Exploration: For this study, we utilized a dataset of credit card transactions obtained from Kaggle's "Credit Card Fraud Detection" dataset [20]. The data consists of financial transactions carried out by cardholders in Europe using credit cards throughout a two-day interval in September 2013. A structured approach was adopted for the analysis, following the process outlined in Figure 1, which guided through the stages of dataset acquisition, preprocessing, model training, evaluation, and comparative analysis. The dataset comprises 284,807 transactions, with about 0.172% identified as fraudulent. This extreme class imbalance represents a significant potential bias that reflects real-world fraud patterns but requires explicit mitigation strategies to prevent models from developing systematic bias toward the majority class. It includes 30 numeric attributes (V1 to V28), Time, and Amount. The last column of the dataset denotes the transaction type (1 for fraudulent transactions and 0 for others), while attributes V1 through V28 remain anonymous for security purposes [21]. The dataset ensures data privacy through PCA transformation of most features (V1-V28), with original transaction details anonymized due to confidentiality requirements, thereby protecting cardholder identities while still enabling effective fraud detection analysis. This dataset has been previously utilized in [21][22].

The initial exploration of the dataset involved loading it and performing a detailed analysis to obtain a thorough comprehension of its attributes. To achieve this, we computed descriptive statistics to analyze the distribution of transaction classes and amounts. To visualize the distribution of transaction classes, we created a bar plot as in Figure 2. Figure 2 displays the transaction class distribution, highlighting the significant imbalance between normal and fraudulent transactions. Normal transactions dominate with approximately 284,000 cases, while fraudulent transactions are barely visible on the chart, representing only about 0.172% of the dataset.

To further investigate the distribution of transaction amounts, histograms were developed as in Figure 3 to create visual representations illustrating the distribution of transaction amounts. These histograms provided a clear understanding of the range of transaction amounts and the frequency of transactions within each range. Finally, the relationship between time and transaction amount for fraudulent transactions was analyzed by creating a scatter plot, as shown in Figure 4. This plot was used to identify trends in the data and determine whether a correlation existed between transaction time and amount.

Data Preprocessing:

Before initiating model training, the dataset underwent preprocessing, including normalization of the 'Amount' feature using the Robust Scaler to reduce the impact of outliers. This normalization technique is robust to outliers, ensuring that extreme values in the 'Amount' feature do not unduly influence the model training process. Robust scaling transforms the data through a process of median removal followed by rescaling following the interquartile range. Additionally, random under-sampling was performed to address class imbalance. To ensure that the models could efficiently learn from the data and generate reliable outputs, these preprocessing steps were crucial.

Robust Scalar:

Machine learning and statistics professionals frequently employ the Robust Scaler data preparation approach to scale numerical features in a dataset. It is particularly advantageous for utilizing datasets that contain outliers or are not normally distributed. By using the Robust Scaler, the scale of the features becomes more consistent, enhancing the effectiveness of some machine learning algorithms particularly those sensitive to feature scaling, as in equation 1 [23]. Equation 1 scales features while minimizing the impact of outliers, ensuring more stable model performance.

$$Scaled\ Value\ = \frac{Original\ Value - Input\ Median}{Input\ IQR} \quad (1)$$

Random Under Sampling:

Class imbalance is a frequent problem in datasets, predominantly in applications like fraud analysis, where deceptive financial activities constitute a statistically minor subset within the broader spectrum of transactions. To tackle this issue, random under-sampling was used to utilize random removal of elements from the preponderant class (non-fraudulent transactions) to foster a more equilibrated representation across classes.

In this study, we employed Random Under sampling to create a more balanced dataset. This process involves the use of a sampling function that randomly selects a specified number of instances from the initial dataset. The focus is primarily on the Majority Class (MC), which contains the majority of samples in the initial dataset. The Number of Samples (NS) parameter is crucial in this method since it outlines the volume of data points included in the resulting undersampled set. Proper adjustment of NS can result in a more balanced representation of classes, vital for constructing machine learning models capable of effectively distinguishing between different classes without bias towards the majority class.

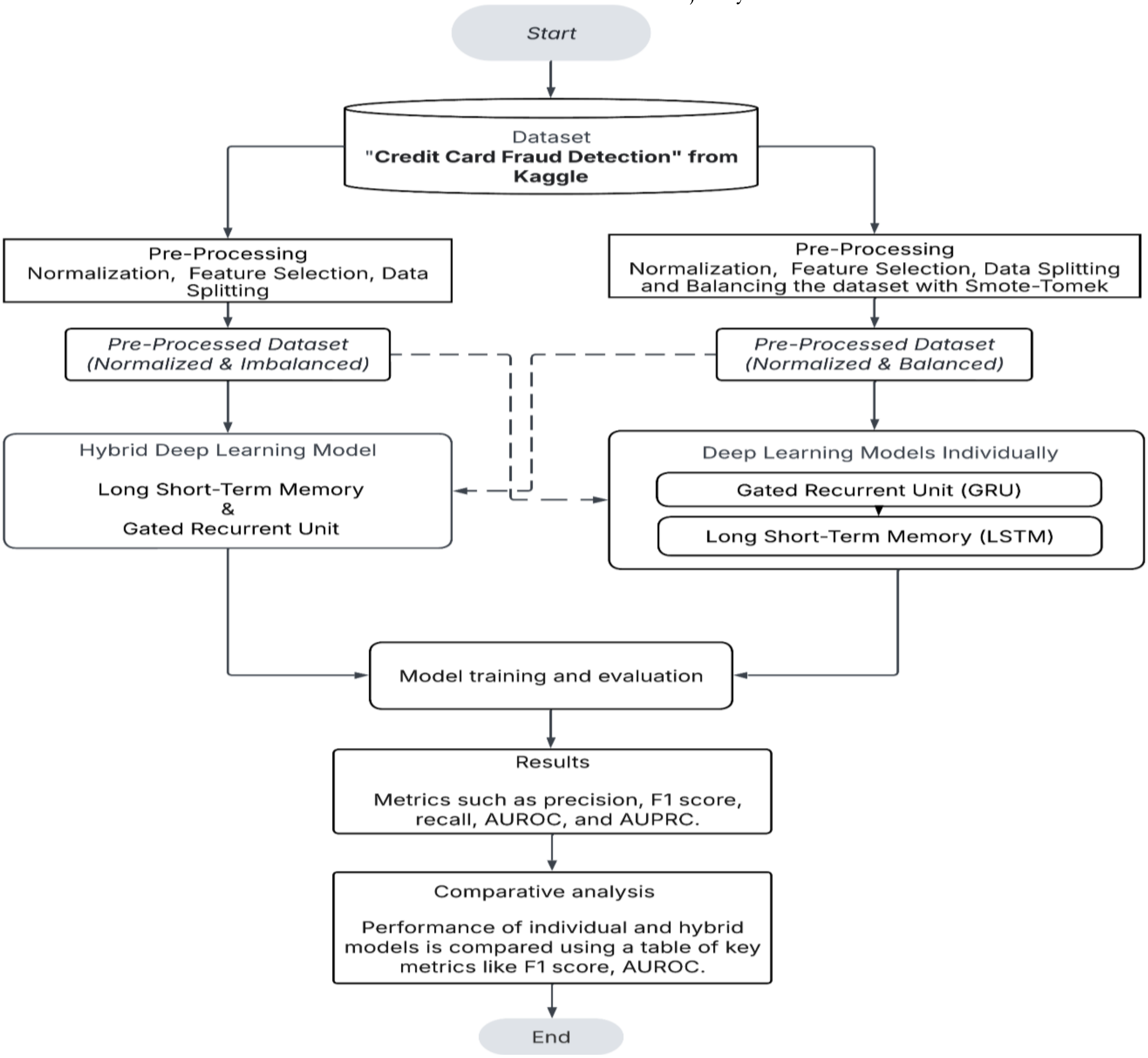
Start
Dataset
"Credit Card Fraud Detection" from Kaggle
Pre-Processing
Normalization, Feature Selection, Data Splitting
Pre-Processing
Normalization, Feature Selection, Data Splitting and Balancing the dataset with Smote-Tomek
Pre-Processed Dataset (Normalized & Imbalanced)
Pre-Processed Dataset (Normalized & Balanced)
Hybrid Deep Learning Model
Long Short-Term Memory
&
Gated Recurrent Unit
Deep Learning Models Individually
Gated Recurrent Unit (GRU)
Long Short-Term Memory (LSTM)
Model training and evaluation
Results
Metrics such as precision, F1 score, recall, AUROC, and AUPRC.
Comparative analysis
Performance of individual and hybrid models is compared using a table of key metrics like F1 score, AUROC.
End

Figure 1. Dataflow of Research Methodology for Addressing Class Imbalance in Fraud Detection

Feature Selection:

Selecting features is crucial for constructing powerful machines or deep learning models. To classify the most informative features, we used the Random Forest algorithm. Random Forest systematically investigates the correlation between feature quantity adjustments and pivotal performance parameters, including the area under the precision-recall curve (AUC-PR), accuracy, F1 score, recall, precision, and the area under the ROC curve (AUC).

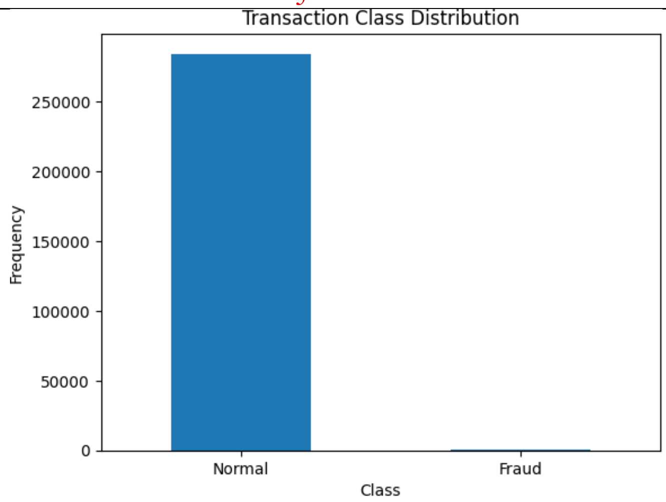

Figure 2. Class Distribution of non-fraud and fraud

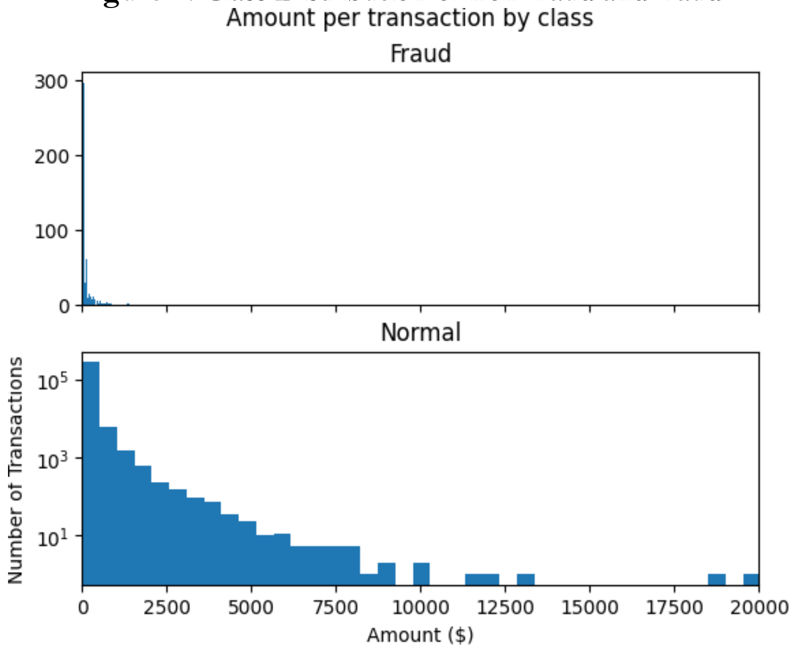

Figure 3. Transaction amount distribution

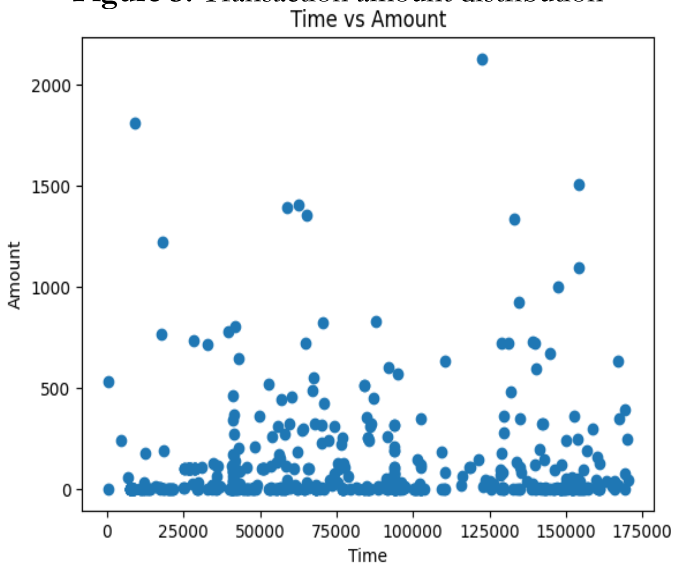

Figure 4. Relationship between time & amount

By employing logistic regression, it evaluates the performance of each feature subset against different thresholds for the number of features. Subsequently, these metrics are plotted against the corresponding threshold values, providing valuable insights into the optimal number of features crucial for our model. This analytical process is presented in Figure 5. The importance of each feature was visualized using a bar plot in Figure 6. Subsequently, a subset of features with the highest importance scores was selected for model training. This step was important to ensure that the models could focus on the most relevant features and avoid overfitting. To further address overfitting, techniques such as dropout and early stopping have been implemented in the model architecture and training process. Smote-Tomek was applied to the dataset after the selection of features.

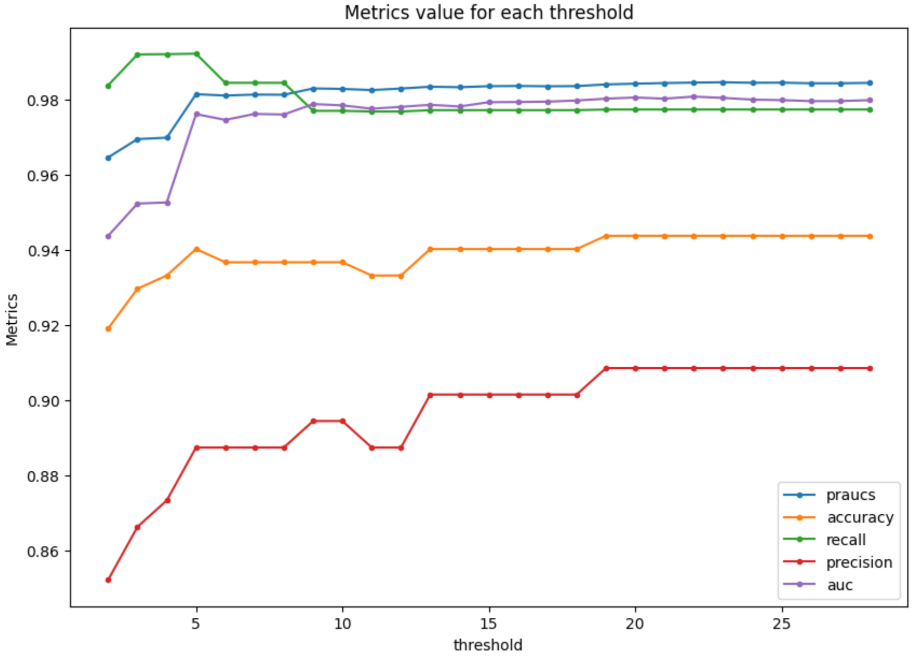

Figure 5. Metric values for each threshold

Random Forest:

The Random Forest method employs ensemble learning, which involves generating multiple decision tree models and synthesizing their predictions to derive a conclusive result. The Random Forest algorithm functions through a series of steps designed to create a robust ensemble of decision trees. Initially, it selects m features randomly from the total set of M features. Using these chosen features, it builds a decision tree by dividing the data into smaller parts based on feature thresholds that minimize impurity. This process was repeated multiple times, resulting in the creation of numerous decision trees that collectively form the forest. The ultimate classification is derived through the aggregation of individual tree outputs within the ensemble framework. For classification tasks, the most frequently occurring prediction across all trees is used as the final output.

Logistic Regression:

A predictive analytics approach used for estimating binary dependent variables predicts the probability that a given input corresponds to a specific category. The functional relationship between explanatory variables and binary response is typically characterized by the logistic function.

Logistic regression forecasts the likelihood of a particular input being part of a particular class via the logistic function 2 – 5 [24]. Equations 2–5 describe logistic regression, which is crucial for fraud detection as it estimates the probability of a transaction being fraudulent.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \ (2)$$

where the linear combination of the input attribute weights is denoted by z:

$$z = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

Here, $x_1, x_2, \ldots, x_n$ are the input features, while $w_0, w_1, w_2, \ldots, w_n$ are the model's coefficients (weights). The logistic regression model predicts the following for a given input's probability of falling into the positive class (y = 1):

$$P(y = 1|x) = \sigma(z) \ (4)$$

And the probability that it belongs to the negative class (y = 0) is:

$$P(y = 0|x) = 1 - \sigma(z)(5)$$

To categorize an input, a threshold (usually 0.5) is selected. If (P(y=1 | x) > 0.5), the input is labeled as part of the positive category; otherwise, it is labeled as part of the negative category.

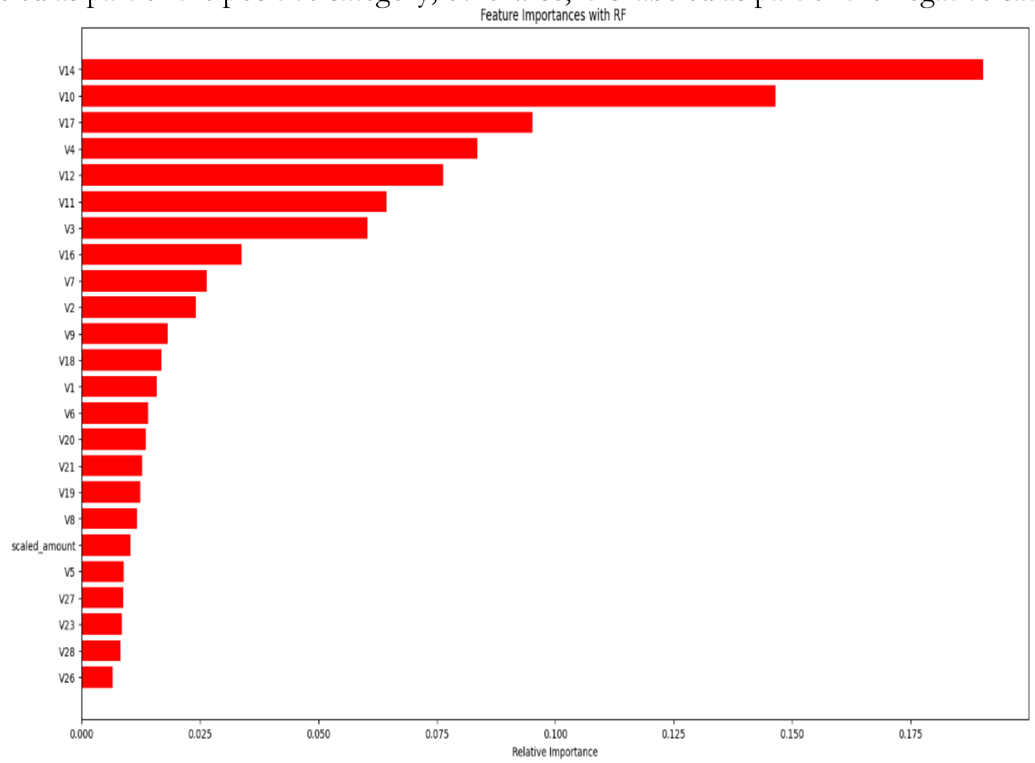

Figure 6. Important Features with Random Forest

Smote-Tomek:

Synthetic Minority Over-sampling Technique, also known as SMOTE, is a methodology employed for handling class imbalances via making artificially created data to augment the minority class. Tomek links denote pairs of instances, where one instance is extracted from the majority class and the other is selected from the minority class, that are close in proximity despite belonging to distinct classes. SMOTE-Tomek combines the oversampling technique of SMOTE with the under-sampling technique of Tomek links to create a more balanced dataset while simultaneously improving the separation between classes.

To address the class imbalance, a combined approach utilizing the Synthetic Minority Over-sampling Technique (SMOTE) and Tomek links was applied, resulting in a balanced dataset referred to as ST. We applied function S to the original dataset O that utilizes SMOTE to generate artificial instances for underrepresented categories. This step increases the representation of underrepresented classes. Subsequently, we employ function TL to identify Tomek links between samples. Tomek links denote proximal instance pairs of opposing classes; eliminating the majority class instance of these links helps to clarify the class boundaries. This two-step process of oversampling followed by targeted under-sampling not only makes the dataset more balanced but also improves the quality of the decision boundary between classes, potentially improving the deep learning models' effectiveness.

Model Training and Evaluation:

Three distinct deep learning models have been trained and assessed, one of them is Long Short-Term Memory (LSTM), another is Gated Recurrent Unit (GRU), as well as an additional hybrid model that incorporates LSTM and GRU layers - to determine their effectiveness in identifying fraudulent transactions. Each model underwent training using the preprocessed dataset and assessment substantiated by a range of performance benchmarks, particularly accuracy, precision, recall, F1 Score, AUROC, and AUPRC. Training histories for each model were visualized using line plots in Figure 8, Figure 11, and Figure 14, while ROC curves and Precision-Recall curves were plotted to assess model differentiating ability. This step was crucial to ensure that the models could effectively detect fraudulent transactions.

Long Short-Term Memory (LSTM):

LSTM has emerged as a sophisticated variant of recurrent neural networks (RNNs), purposefully engineered to circumvent the traditional RNNs' drawbacks in capturing and maintaining long-range contextual information. It employs specialized gates and memory cells to modulate information flow, thus enhancing its capacity to preserve relevant data over prolonged sequences. Hochreiter and Schmidhuber initially introduced LSTM. Key equations governing LSTM are given below 6 - 13 [23].

$$\tan h(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \ (6)$$

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \ (7)$$

$$i_t = \sigma\left(w_{xi}^T x_{(t)} + w_{hi}^T h_{(t-1)} + b_i\right)(8)$$

$$f_t = \sigma\left(w_{fx}^T x_{(t)} + w_{hf}^T h_{(t-1)} + b_f\right)(9)$$

$$o_t = \sigma\left(w_{xo}^T x_{(t)} + w_{ho}^T h_{(t-1)} + b_o\right)(10)$$

$$g_t = \tanh\left(w_{xg}^T x_{(t)} + w_{hg}^T h_{(t-1)} + b_g\right)(11)$$

$$c_t = f_t \odot c_{(t-1)} + i_t \odot g_t (12)$$

$$h_t = o_t \odot \tanh c_{(t)} \ (13)$$

Equations 6–13 define the Long Short-Term Memory model, which is crucial for detecting fraud patterns in sequential transaction data. Long Short-Term Memory architecture incorporates several key components that work in concert to process sequential data effectively. New information gets integrated into the cell state through the input gate. i_t, while the forget gate f_t Discards unnecessary information from the prior cell state in a selective manner. The memory component is the cell state. c_t, which retains data over long sequences. The flow of learned features from the cell's memory to the hidden layer representation is under the output gate's o_t Control, determining how the internal memory influences the output. The hidden state h_t Encapsulates the network's understanding of the input sequence up to the current time step. The network processes an input vector. x_t At every time step t, represents the current sequence element's features. The hidden state h_{t-1} and cell state c_{t-1} from the previous time step carries forward relevant information from prior inputs, enabling the network to maintain context over long sequences. Bias vectors b and weight matrices W are used in the various computations within the LSTM cell, allowing the network to understand and adjust to the specific patterns found in the input data.

Gated Recurrent Unit (GRU):

GRU is a refined RNN structure developed specifically to address the problem of gradients vanishing during backpropagation. This approach refines the LSTM model by consolidating the input and forget gates into a single update gate, thereby improving computational efficiency. The innovative GRU was conceptualized and presented to the field of deep learning by Cho et al. [25]. Key equations governing GRU are in 14 - 16:

$$\mathrm{r_t} = \sigma\left(\mathrm{w_{xr}^T x_{(t)}} + \mathrm{w_{hr}^T o_{(t-1)}} + \mathrm{b_r}\right)(14)$$

$$\mathrm{z_t} = \sigma\left(\mathrm{w_{xz}^T x_{(t)}} + \mathrm{w_o^T z\, o_{(t-1)}} + \mathrm{b_z}\right)(15)$$

$$\mathrm{o_t} = \mathrm{z_t} \odot \mathrm{o_{t-1}} + (1 - \mathrm{z_t}) \odot \tilde{\mathrm{o}}_\mathrm{t}(16)$$

GRU improves fraud detection by capturing temporal patterns in transaction sequences while being more computationally efficient than LSTM, as shown in Equations 14–16. In the context of Gated Recurrent Units (GRUs), several key components work in tandem to process sequential data. The network gets an input. x_t At every time step t, using the preceding time step's concealed state. $o_{(t-1)}$. The significance of the update gate z_t Lies in its role in deciding the extent to which the previous state should be preserved. The model employs various weight matrices. $w_{xr}, w_{hr}, w_{xz}, w_o$ and bias vectors b_r, b_z To perform its computations. The candidate activation o_t Represents a potential new hidden state. The sigmoid activation function σ is used to compute gate values, ensuring they fall between 0 and 1. Element-wise multiplication, denoted by $\odot$ is utilized in several operations within the GRU, allowing for fine-grained control over information flow. This architecture enables GRUs to effectively capture and propagate pertinent data across extended sequences while addressing the issue of the diminishing gradient.

Hybrid Model (LSTM + GRU):

The hybrid model integrates the advantages of both LSTM and GRU architectures through the use of alternating LSTM and GRU layers. The hybrid approach aims to capitalize on LSTM's robust memory retention capabilities and GRU's enhanced computational efficiency. Each layer processes the input sequences bidirectionally, capturing information from both past and future time steps.

Proposed LSTM-GRU Hybrid Model:

The neural network architecture begins with the initialization of a Sequential model. The first layer is a 100-unit bidirectional LSTM, configured to return sequences and shaped to match the input dimensions of the training data. Following this, a dropout regularization strategy is employed with a 0.3 dropout probability to prevent overfitting. The next layer comprises a Bidirectional GRU with 50 units, also returning sequences, followed by another Dropout layer with a 0.3 probability. The network then incorporates a densely interconnected layer of 64 artificial neurons employing RELU activation. The last layer consists of a dense layer with just one neuron and uses sigmoid activation. Adam is the model optimizer with a rate of learning 1×10^{-4} (0.0001). After compilation, the model's architecture is displayed using a summary function. The training process is subsequently initiated, running for 20 epochs with a batch size of 64 samples and incorporating specified validation data to track performance throughout the training phase.

Evaluation Metrics:

Accuracy: In the domain of classification model evaluation, accuracy is a commonly applied metric. It is ascertained by computing the proportion of accurately classified samples relative to the complete set of observations as in 17 [26].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}(17)$$

In the evaluation of classification models, four fundamental metrics are employed to assess performance comprehensively. True Positives (TP) represent the number of instances accurately classified to be in a positive class, signifying the algorithm's capacity to precisely detect true positive occurrences. Conversely, TNs, or true negatives, indicate the count of cases that are appropriately categorized as members of the negative class, showcasing the model's proficiency in identifying negative cases. False Positives (FP) quantify the instances categorized as positive despite their actual negative classification, highlighting potential over-sensitivity in the model. Lastly, False Negatives (FN) enumerate the frequency of samples incorrectly labeled as negative despite being truly positive, indicating potential under sensitivity. These metrics form the cornerstone of various performance measures such as accuracy, precision, recall, and F1-score, offering a sophisticated comprehension of a model's advantages and weaknesses across different aspects of classification.

F1 Score:

It is a composite metric that optimizes the tradeoff between recall and precision, calculated as the harmonic average of these two indicators, offering a nuanced assessment of a model's predictive efficacy, provides a unified metric that considers both aspects equally as in 18 [26].

$$\mathrm{F1} = 2 \times \frac{\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} (18)$$

Recall:

Recall, interchangeably used with sensitivity and accurate measurement of the proportion of true positives. It is determined as in 19 [26].

$$\mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} (19)$$

Precision:

The precision score assesses the reliability of positive classifications, computed as a quotient of actually true positives to the sum of all cases the model designated as positive. It is determined as in 20 [26].

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} (20)$$

Area Under the ROC Curve (AUROC):

AUROC functions as a performance metric for assessing a binary classifier's discriminative capacity between positive and negative classes across varying decision thresholds. This scalar value is derived by calculating the area subtended by the Receiver Operating Characteristic curve (ROC), which graphically represents interdependence between the false positive and true positive rate at diverse classification thresholds. AUROC values are constrained within the interval [0,1], with higher magnitudes signifying enhanced model efficacy in class separation.

Area Under the Precision-Recall Curve (AUPRC):

AUPRC serves as a crucial measure for assessing the performance of models for binary classification, particularly when handling unbalanced class distributions. It evaluates the precision-recall curve area, illustrating the precision plotted against the recall for various threshold values. The AUPRC is confined to the range of 0 to 1, the greater score signifies superior model performance, particularly in situations where the positive class (e.g., fraudulent transactions) is rare.

These preprocessing techniques and evaluation metrics were instrumental in making certain that the models were trained using a balanced and adequate dataset and evaluated using appropriate performance metrics that account for class imbalance and the costs associated with misclassification.

Results and Discussion:

Model Performance Evaluation:

LSTM Model:

After training with SMOTE-Tomek to address the class imbalance, the LSTM model achieved 87.37% precision, 86.46% recall, and 98.92% accuracy. This model demonstrated a balanced performance in identifying fraudulent transactions, as evidenced by an F1 Score of

86.91%. The classifier's effectiveness in distinguishing between fraudulent and non-fraudulent transactions is demonstrated by the AUROC score of 97.46% and AUPRC score of 86.63% as represented in Figure 7. However, when the LSTM model was trained without applying SMOTE-Tomek, it achieved a slightly higher precision of 90.11%, recall of 85.42%, and accuracy of 99.58%. This led to a significantly higher F1 Score of 87.70%, while the AUROC and AUPRC scores decreased to 97.99% and 78.71%, respectively.

Figure 8 presents the training and validation AUC over epochs. The training AUC remains consistently close to 1.0, while the validation AUC fluctuates slightly between 0.95 and 1.0. This indicates that, although the model generalizes well, there are signs of mild overfitting, as the training performance is slightly higher than the validation performance. However, the gap is minimal, implying that the overfitting is not severe, and the model continues to exhibit strong generalization capabilities. The confusion matrix displayed in the LSTM model provides a visual representation of its comprehensive classification performance as illustrated in Figure 9.

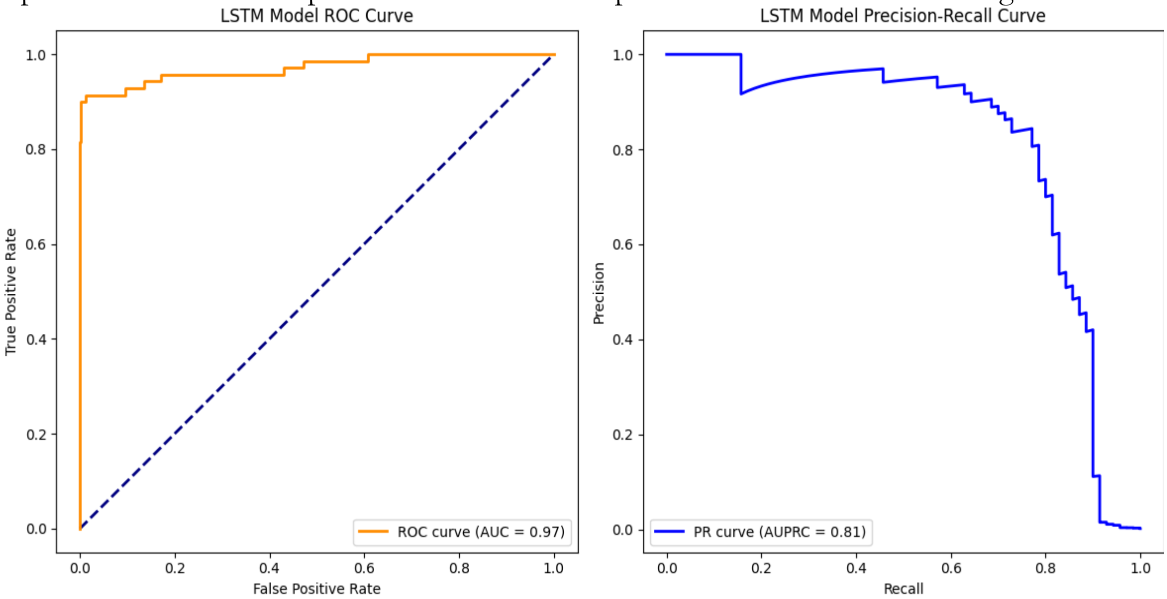

Figure 7. AUROC and AUPRC for LSTM model with Smote-Tomek

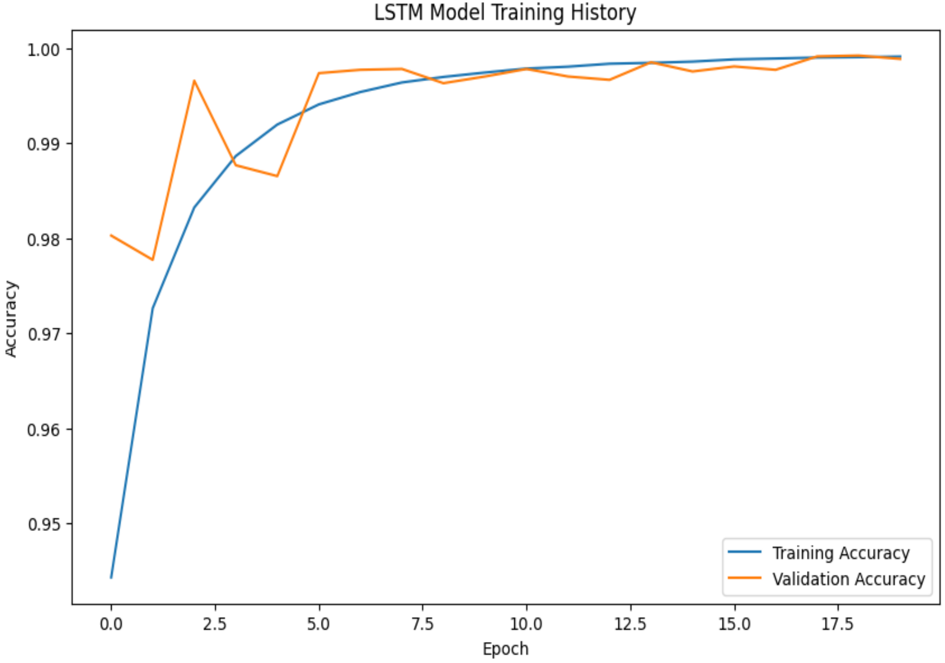

Figure 8. Long Short-Term Memory (LSTM) Training History

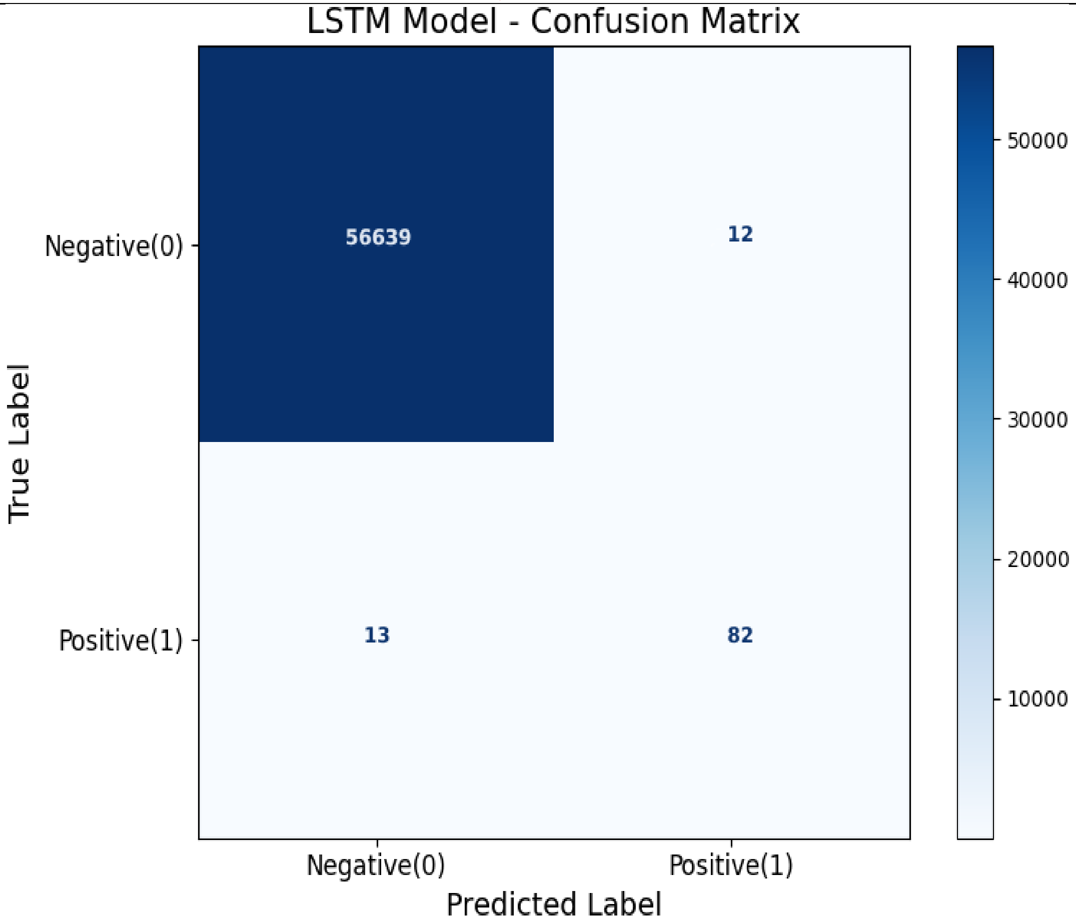

Figure 9. Confusion Matrix for LSTM model with Smote-Tomek

GRU Model:

The GRU model, trained with SMOTE-Tomek to address class imbalance, exhibited excellent performance metrics, boasting a 98.97% accuracy. It demonstrated an F1 Score of 86.49% with a precision of 89.89% and a recall of 83.33%. The GRU model also demonstrated superior differentiating ability, as reflected by an AUROC score of 95.82% and an AUPRC score of 80.01% as in Figure 10. In contrast, when trained without SMOTE-Tomek, the GRU model obtained 97.79% accuracy, 82.76% precision, and 75.00% recall. The F1 Score was 78.69%, and the AUROC and AUPRC scores were 94.91% and 72.69%, respectively. This demonstrates that using SMOTE-Tomek significantly improved the effectiveness of the model in identifying fraudulent transactions.

Figure 11 presents the GRU training and validation history. The training AUC starts at 0.98 and reaches 1.0, while the validation AUC slightly decreases from 0.98 to 0.97. This suggests that the model is learning effectively but shows a slight tendency toward overfitting, as the training performance continues to improve while the validation performance experiences a minor decline. However, the validation AUC remains high, and the model still generalizes well, indicating minimal overfitting. The model's classification performance is comprehensively visualized in the confusion matrix presented in Figure 12, w hich illustrates the distribution of false positives, true negatives, true positives, and false negatives for the GRU model, along with an in-depth analysis of its prediction accuracy.

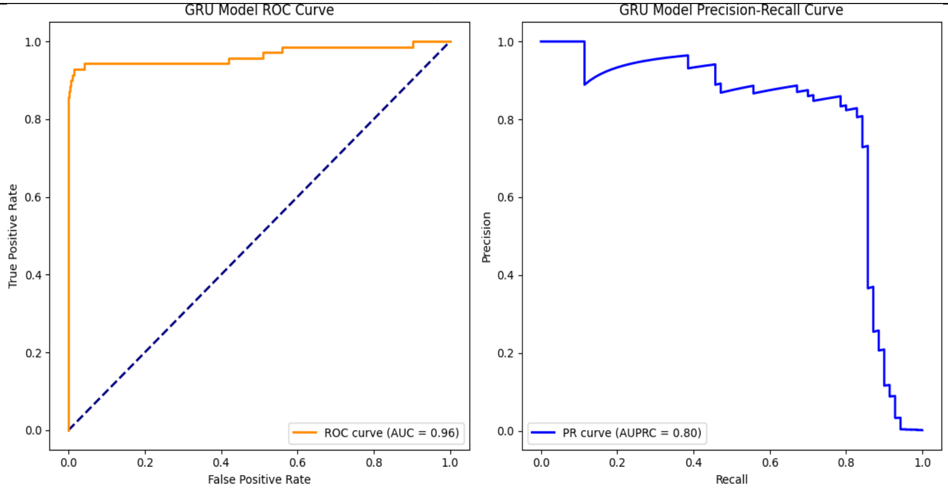

Figure 10. AUROC and AUPRC for GRU model with Smote-Tomek

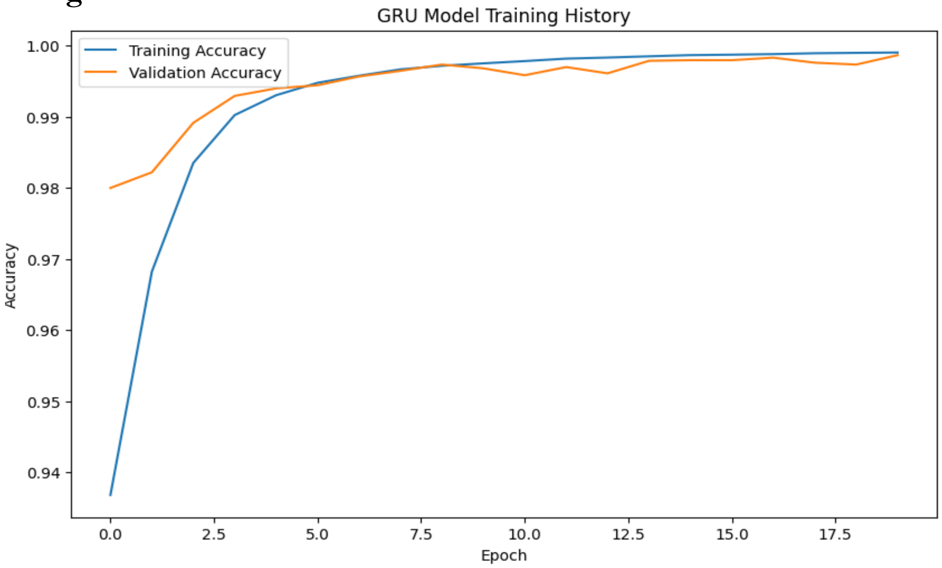

Figure 11. Gated Recurrent Unit (GRU) Training History

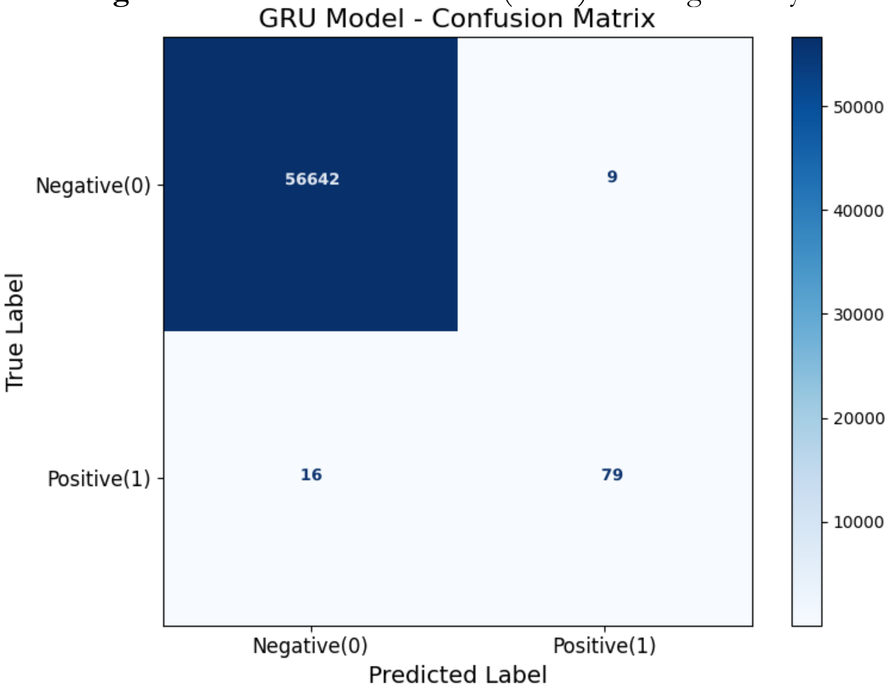

Figure 12. Confusion Matrix for GRU model with Smote-Tomek

Hybrid Model:

The Hybrid model, comprising LSTM and GRU layers, attained an accuracy of 99.56% with SMOTE-Tomek applied. Notably, it achieved the highest precision and recall rates among all models, with values of 91.21% and 86.46% respectively. The Hybrid model yielded an AUROC score of 97.62%, an AUPRC score of 88.61%, and an F1 Score of 88.77% as shown in Figure 13. In comparison, when trained without SMOTE-Tomek, the Hybrid model obtained 98.16% accuracy, 89.01% precision, and 84.38% recall. The F1 Score was 86.63%, and the AUROC and AUPRC scores were 93.87% and 69.55%, respectively. This comparison underscores the substantial performance improvement when SMOTE-Tomek is utilized.

Figure 14 presents the training and validation history for the Hybrid model. The training AUC increases from 0.99 to 1.0, indicating that the model is learning effectively. However, the validation AUC declines from 0.96 to 0.94, suggesting potential overfitting. This occurs because while the model continues to improve on the training set, its performance on unseen data slightly decreases. The widening gap between training and validation AUC indicates that the model may be memorizing training patterns rather than generalizing well, though the overall validation performance remains high. The model's classification performance is further illustrated in the confusion matrix presented in Figure 15.

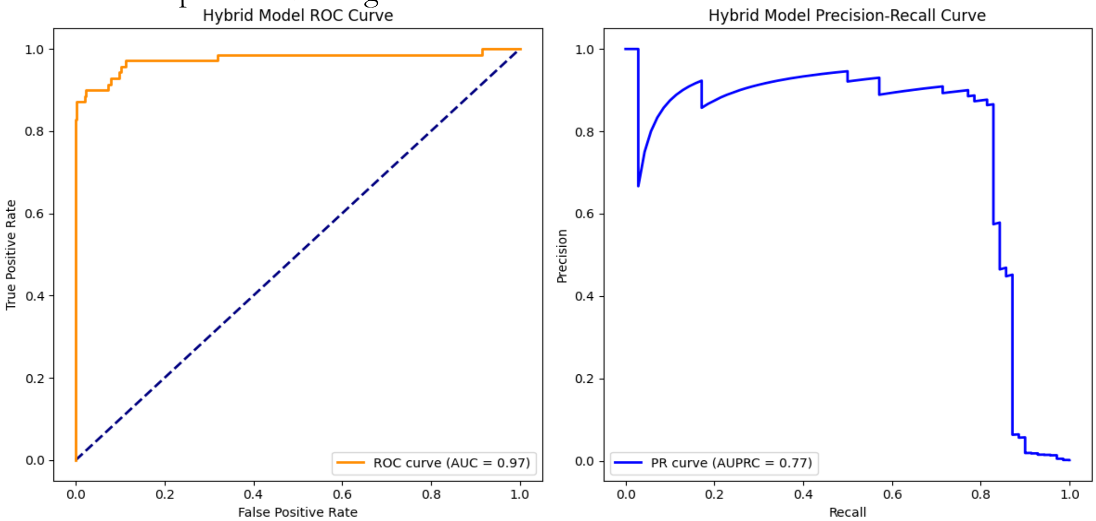

Figure 13. AUROC and AUPRC for Hybrid model with Smote-Tomek

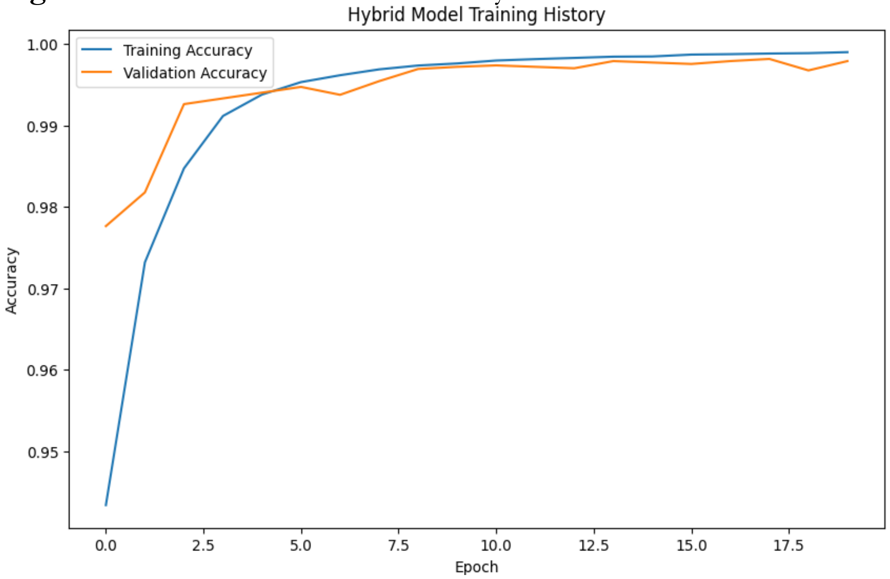

Figure 14. Hybrid Model Training History

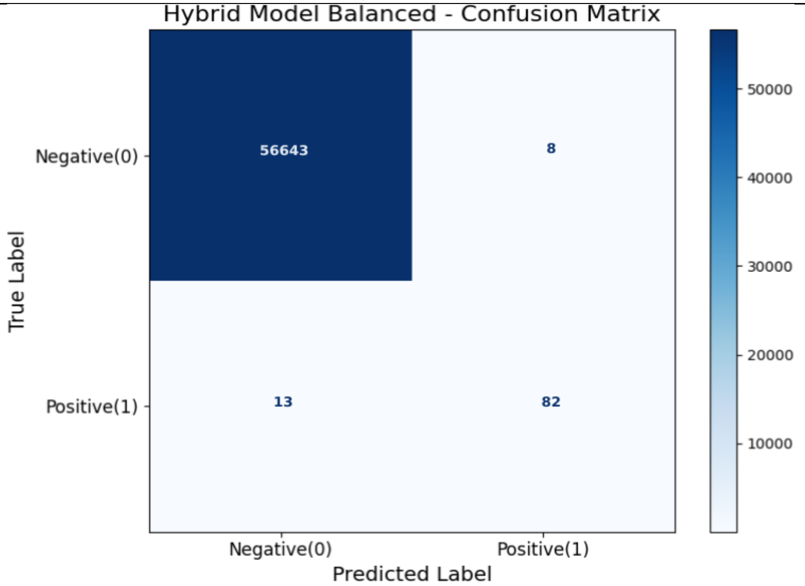

Figure 15. Confusion Matrix for Hybrid model with Smote-Tomek

The findings highlight the effectiveness of deep learning models in detecting fraudulent credit card transactions. However, it is essential to acknowledge the significant impact of the SMOTE-TOMEK technique on enhancing the models' performance. Both LSTM and GRU models demonstrated robust performance, with the Hybrid model exhibiting promising results when combined with SMOTE-TOMEK, particularly in terms of precision, AUROC as presented in Table 2.

The suggested Hybrid LSTM-GRU model is compared with cutting-edge models for credit card fraud detection in Table 3, with an emphasis on important performance indicators including accuracy, precision, recall, F1-score, and AUROC.

Table 1. Performance Metrics of Models without SMOTE-TOMEK

Model	Accuracy	Precision	Recall	F1 Score	AUROC	ARC
LSTM	0.995771	0.9011	0.8542	0.8770	0.979885	0.787106
GRU	0.977937	0.8276	0.7500	0.7869	0.949133	0.726898
Hybrid	0.981620	0.8901	0.8438	0.8663	0.938704	0.695542

Table 2. Performance Metrics of Models with SMOTE-TOMEK

Model	Accuracy	Precision	Recall	F1 Score	AUROC	ARC
LSTM	0.989215	0.8737	0.8646	0.8691	0.974551	0.866329
GRU	0.989691	0.8989	0.8333	0.8649	0.958174	0.800134
Hybrid	0.995577	0.9121	0.8646	0.8877	0.976205	0.886068

Table 3. Performance comparison of the proposed approach with other models

Model	Accuracy	Precision	Recall	F1 Score	AUROC	ARC
Hybrid Model (SMOTE-TOMEK)	**0.9956**	**0.9121**	**0.8646**	**0.8877**	**0.9762**	**0.8861**
LSTM (SMOTE-TOMEK)	0.9892	0.8737	0.8646	0.8691	0.9745	0.8663
GRU (SMOTE-TOMEK)	0.9896	0.8989	0.8333	0.8649	0.9582	0.8001
GA-ANN [21]	0.8893	0.8240	0.7876	0.8054	0.9400	-
Adaboost+LGBM Hybrid [7]	-	0.9700	0.6400	0.7700	0.8200	-

Adaboost+XGBOOST [7]]	-	0.9400	0.5900	0.7300	0.7900	-
Optimized LightGBM (O-LightGBM) [8]	0.9840	0.9730	0.4060	0.5690	0.9090	-
LightGBM [8]	0.9992	0.7530	0.7990	0.7690	0.9470	-
XGBoost [8]	0.9993	0.7900	0.8010	0.7900	0.9520	-

The LSTM and GRU models showed balanced performance across multiple metrics, making them suitable choices for practical uses where both precision and recall are essential. On the other hand, the Hybrid model's high precision suggests its potential for use cases prioritizing the minimization of false positives. The AUROC scores illustrate the models' ability to distinguish between fraudulent and non-fraudulent transactions, where higher values reflect superior performance. The observed AUROC scores for all models suggested satisfactory differentiating ability, underscoring their utility in fraud detection tasks.

Overall, the results highlighted the significance of leveraging deep learning techniques for the detection of unauthorized credit card usage. Future studies could explore further model refinement and feature engineering techniques to enhance detection accuracy and robustness. Additionally, real-world deployment considerations, such as computational efficiency and interpretability, warrant further investigation to ensure practical applicability in financial systems.

Conclusion:

In conclusion, this academic exploration strives to address the compelling issue of credit card fraud detection by leveraging advanced deep learning and hybrid algorithms, in combination with data preprocessing methods like SMOTE-Tomek. The study demonstrates the efficacy of a Gated Recurrent Unit, Long Short-Term Memory, and a Hybrid model integrating GRU and LSTM layers in accurately identifying fraudulent transactions within credit card data. Using careful data, preprocessing, feature selection, and model training, this study highlights the effectiveness of deep learning architectures in discerning fraudulent activities amidst legitimate transactions. The evaluation of model performance utilizing performance metrics such as F1 Score, AUROC, recall, accuracy, precision, and AUPRC underscore the robustness and differentiative ability of the proposed models.

Notably, the Hybrid model shows encouraging outcomes, particularly about F1 score (0.8877), recall (0.8646), and precision (0.9121), AUROC (0.976205), and AUPRC (0.886068), showcasing its real-world applicability potential where minimizing false positives is paramount. The results highlight the significance of employing sophisticated deep learning techniques in conjunction with appropriate preprocessing strategies to reduce financial losses and improve the accuracy of fraud detection.

Furthermore, the insights gained from this research provide valuable guidance to financial institutions and stakeholders in selecting and deploying effective fraud detection systems. By continuously refining and optimizing deep learning models, alongside exploring novel preprocessing techniques, the financial industry can bolster its defenses against evolving fraudulent activity involving credit card transactions.

Considering the ever-changing landscape of financial transactions and cyber threats, future research endeavors should focus on further enhancing model robustness, scalability, and interpretability. Additionally, investigations into real-world deployment considerations, computational efficiency, and regulatory compliance are imperative to ensure the practical applicability and efficacy of fraud detection systems in safeguarding financial assets and maintaining trust in digital transactions.

Author's Contribution:

To this work, each author has contributed equally.

Conflict of interest:

No conflicts of interest exist.

References:

[1] Caitlin Mullen, “Card industry’s fraud-fighting efforts pay off: Nilson Report,” *Payments Dive*, 2023, [Online]. Available: https://www.paymentsdive.com/news/card-industry-fraud-fighting-efforts-pay-off-nilson-report-credit-debit/639675/

[2] A. S. Rathore, A. Kumar, D. Tomar, V. Goyal, K. Sarda, and D. Vij, “Credit Card Fraud Detection using Machine Learning,” *Proc. 2021 10th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2021*, pp. 167–171, 2021, doi: 10.1109/SMART52563.2021.9676262.

[3] K. S. and R. G. C. Phua, V. Lee, “A comprehensive survey of data mining-based fraud detection research,” *Artif. Intell.*, 2010, [Online]. Available: https://www.researchgate.net/publication/46887451_A_Comprehensive_Survey_of_Data_Mining-based_Fraud_Detection_Research

[4] R. J. Bolton and D. J. Hand, “Statistical fraud detection: A review,” *Stat. Sci.*, vol. 17, no. 3, pp. 235–255, 2002, [Online]. Available: https://projecteuclid.org/journals/statistical-science/volume-17/issue-3/Statistical-Fraud-Detection-A-Review/10.1214/ss/1042727940.full

[5] J. K. and A. K. S. P. Tiwari, S. Mehta, N. Sakhuja, “Credit card fraud detection using machine learning: a study,” *arXiv Prepr. arXiv2108*, 2021.

[6] M. H. K. Ruixing Ming, “Comparing SMOTE Family Techniques in Predicting Insurance Premium Defaulting using Machine Learning Models,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 9, 2021, doi: 10.14569/IJACSA.2021.0120970.

[7] X. C. Esraa Faisal Malik, Khai Wah Khaw, Bahari Belatonm, Wai Peng Wong, “Credit Card Fraud Detection Using a New Hybrid Machine Learning Architecture,” *Mathematics*, vol. 10, no. 9, p. 1480, 2022, doi: https://doi.org/10.3390/math10091480.

[8] S. L. M. and S. G. S. K. Hashemi, “Fraud Detection in Banking Data by Machine Learning Techniques,” *IEEE Access*, vol. 11, pp. 3034–3043, 2023, doi: 10.1109/ACCESS.2022.3232287.

[9] A. N. Ahmed and R. Saini, “A Survey on Detection of Fraudulent Credit Card Transactions Using Machine Learning Algorithms,” *2023 3rd Int. Conf. Intell. Commun. Comput. Tech. ICCT 2023*, 2023, doi: 10.1109/ICCT56969.2023.10076122.

[10] E. Jayanthi *et al.*, “Cybersecurity enhancement to detect credit card frauds in health care using new machine learning strategies,” *Soft Comput.*, vol. 27, no. 11, pp. 7555–7565, Jun. 2023, doi: 10.1007/S00500-023-07954-Y/METRICS.

[11] I. D. Mienye and Y. Sun, “A Deep Learning Ensemble With Data Resampling for Credit Card Fraud Detection,” *IEEE Access*, vol. 11, pp. 30628–30638, 2023, doi: 10.1109/ACCESS.2023.3262020.

[12] K. I. Alkhatib, A. I. Al-Aiad, M. H. Almahmoud, and O. N. Elayan, “Credit Card Fraud Detection Based on Deep Neural Network Approach,” *2021 12th Int. Conf. Inf. Commun. Syst. ICICS 2021*, pp. 153–156, May 2021, doi: 10.1109/ICICS52457.2021.9464555.

[13] S. Kumar, V. K. Gunjan, M. D. Ansari, and R. Pathak, “Credit Card Fraud Detection Using Support Vector Machine,” *Lect. Notes Networks Syst.*, vol. 237, pp. 27–37, 2022, doi: 10.1007/978-981-16-6407-6_3.

[14] J. Karthika and A. Senthilselvi, “Credit Card Fraud Detection based on Ensemble Machine Learning Classifiers,” *3rd Int. Conf. Electron. Sustain. Commun. Syst. ICESC 2022 - Proc.*, pp. 1604–1610, 2022, doi: 10.1109/ICESC54411.2022.9885649.

[15] M. R. and M. A. F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, “Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms,” *IEEE Access*, vol. 10, pp. 39700–39715, 2022, doi: 10.1109/ACCESS.2022.3166891.

[16] E. Strelcenia and S. Prakoonwit, “A New GAN-based data augmentation method for

Handling Class Imbalance in Credit Card Fraud detection," *Proc. 10th Int. Conf. Signal Process. Integr. Networks, SPIN 2023*, pp. 627–634, 2023, doi: 10.1109/SPIN57001.2023.10116543.

[17] Amerah Alabrah, "An Improved CCF Detector to Handle the Problem of Class Imbalance with Outlier Normalization Using IQR Method," *Sensors*, vol. 23, no. 9, p. 4406, 2023, doi: https://doi.org/10.3390/s23094406.

[18] V. S. B. and R. J. A. Mahajan, "Credit Card Fraud Detection using Logistic Regression with Imbalanced Dataset," *10th Int. Conf. Comput. Sustain. Glob. Dev. (INDIACom), New Delhi, India*, pp. 339–342, 2023, [Online]. Available: https://ieeexplore.ieee.org/document/10112302

[19] A. S. Alexey Ruchay, Elena Feldman, Dmitriy Cherbadzhi, "The Imbalanced Classification of Fraudulent Bank Transactions Using Machine Learning," *Mathematics*, vol. 11, no. 13, p. 2862, 2023, doi: https://doi.org/10.3390/math11132862.

[20] MACHINE LEARNING GROUP - ULB, "Credit Card Fraud Detection," *Kaggle*, 2017.

[21] Y. S. & Z. W. Emmanuel Ileberi, "A machine learning based credit card fraud detection using the GA algorithm for feature selection," *J. Big Data*, vol. 9, no. 24, 2022, doi: https://doi.org/10.1186/s40537-022-00573-8.

[22] B. E. O. & J. J. Ibtissam Benchaji, Samira Douzi, "Enhanced credit card fraud detection based on attention mechanism and LSTM deep model," *J. Big Data*, vol. 8, no. 151, 2021, doi: https://doi.org/10.1186/s40537-021-00541-8.

[23] Y. Singh, "Robust Scaling: Why and How to Use It to Handle Outliers," *Proclus Acad.*, 2022, [Online]. Available: https://proclusacademy.com/blog/robust-scaler-outliers/

[24] Google Developers, "Calculating a Probability - Logistic Regression," *Google Dev.*, 2023, [Online]. Available: https://developers.google.com/machine-learning/crash-course/logistic-regression/sigmoid-function

[25] Y. B. Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," *Assoc. Comput. Linguist.*, pp. 1724–1734, 2014, doi: 10.3115/v1/D14-1179.

[26] S. S. and N. H. A. H. M. O. S. Yee, "Credit card fraud detection using machine learning as data mining technique," *Seybold Rep.*, vol. 15, no. 9, pp. 2431–2436, 2020, [Online]. Available: https://www.researchgate.net/publication/344788652_CREDIT_CARD_FRAUD_DETECTION_USING_DATA_MINING_TECHNIQUES