***Microsoft Machine Learning Graduation project***

***Group name:*** *FIN AI*

***Project name:*** *Fraud Detection in Financial Transactions*

***Objective:***  *Designing a fraud detection model*

***Members:***

*Aya Emad, Hossam Eldin Mohamed, Mostafa Maged, Omar Ashraf, Thorya Mahfouz*

*A credit card with different symbols

Description automatically generated*

Contents

[**Project Title:** 3](#_heading=h.gjdgxs)

[**Introduction:** 3](#_heading=h.30j0zll)

[Objective: 3](#_heading=h.1fob9te)

[**Methodology:** 3](#_heading=h.3znysh7)

[∙](#_heading=h.2et92p0) Python & Jupyter Notebooks: 3

[∙](#_heading=h.tyjcwt) MLflow: 3

[∙](#_heading=h.3dy6vkm) XGBoost, RandomForest, LightGBM: 3

[∙](#_heading=h.1t3h5sf) PyTorch: 3

[∙](#_heading=h.4d34og8) Category Encoders & Scikit-learn 3

[∙](#_heading=h.2s8eyo1) Pyngrok: 3

[Approach: 3](#_heading=h.17dp8vu)

[1.](#_heading=h.3rdcrjn) Data Collection and Preprocessing: 3

2[.](#_heading=h.lnxbz9) Model Training and Evaluation: 4

3[.](#_heading=h.26in1rg) Data Augmentation with TCGANs: 4

[4.](#_heading=h.35nkun2) Deployment and Monitoring: 4

[**Results:** 5](#_heading=h.1ksv4uv)

[Demonstration: 5](#_heading=h.44sinio)

[**Challenges:** 8](#_heading=h.z337ya)

[How They Were Overcome: 9](#_heading=h.3j2qqm3)

[**Future Work:** 9](#_heading=h.1y810tw)

[Next Steps: 9](#_heading=h.4i7ojhp)

[**Conclusion:** 9](#_heading=h.2xcytpi)

**Project Title:**  
Fraud Detection in Financial Transactions Using MLOps and GANs

**Introduction:**  
Problem Statement:  
When it comes to digital transactions, fraud has become an increasingly serious issue. With online payments on the rise, both consumers and financial institutions face significant risks from fraudulent activities. This project aims to tackle this challenge head-on by developing a machine learning model designed to detect fraudulent transactions effectively. The innovative twist? We’re using synthetic data generated by Generative Adversarial Networks (GANs) to enhance our model's training process. This approach not only improves the model's accuracy but also ensures it can adapt to a wide range of fraud scenarios.

Objective:  
The key objective is to design a fraud detection model that leverages advanced machine learning techniques, including GANs for data augmentation, to improve detection accuracy. This involves implementing a robust pipeline for model development, training, and deployment, incorporating continuous tracking and version control of models and experiments through MLflow.

**Methodology:**  
Tools & Technologies:

* Python & Jupyter Notebooks: Used for developing and running the project code, leveraging the extensive ecosystem of data science libraries.
* MLflow: Employed for managing and tracking experiments, facilitating model versioning, and ensuring reproducibility.
* XGBoost, RandomForest, LightGBM: Utilized for their efficiency in handling imbalanced datasets and high-dimensional data, crucial for fraud detection tasks.
* PyTorch: Used to build GANs for generating synthetic transaction data.
* Category Encoders & Scikit-learn**:** Used for data preprocessing, including encoding categorical variables and standardizing numerical features.
* Pyngrok: Deployed to create a secure tunnel to host the MLflow UI, enabling remote access for tracking experiments.

## Approach:

### Data Collection and Preprocessing:

* + **Data Loading and Initial Cleaning:** Data was sourced from Kaggle, split into training and testing sets, and concatenated for comprehensive analysis. Missing values were imputed, and unnecessary columns were dropped to maintain a clean dataset.
  + **Feature Engineering:** Advanced feature engineering was performed, including extracting time-based features (e.g., hour of transaction), calculating age from date of birth, and measuring geographical displacement between customer and merchant locations. These features were categorized into bins to enhance model interpretability and performance.
  + **Exploratory Data Analysis (EDA):** Data visualization techniques were employed to understand the distribution of transaction amounts and identify potential outliers. A box plot of transaction amounts highlighted the quartiles, which are crucial in understanding the typical transaction ranges and identifying anomalies.

### Model Training and Evaluation:

* + **Model Selection and Tuning:** Various classification models, including Logistic Regression, Random Forest, SVM, XGBoost, and LightGBM, were evaluated. The models were fitted using a pipeline approach that integrates preprocessing steps, ensuring data consistency throughout the training process.
  + **Experiment Tracking with MLflow:** Each model's performance was tracked using MLflow, logging parameters, metrics, and model versions. This setup ensured systematic experimentation and allowed for easy comparisons between models.
  + **Feature Importance Analysis:** Random Forest was used to assess feature importance, revealing that transaction amount, category, job, state, and displacement were significant predictors of fraud. This insight guided further model refinement.

### Data Augmentation with GANs:

**GAN Implementation:** A TCGAN was constructed with a generator and discriminator network to produce realistic synthetic transaction data. The generator creates fake samples while the discriminator evaluates their authenticity. This iterative process enhances the generator's ability to create indistinguishable synthetic data from real transactions.

Conditional Tabular GAN specifically designed for generating synthetic tabular data. It introduces key innovations:

o Conditional Generation: By conditioning the generation process on specific categories, CTGAN can better handle categorical variables and their imbalances.

o Mode-Specific Normalization: the model learns multimodal distributions. which are common in real-world datasets.

o Training Strategy: CTGAN uses a specialized **training-by-sampling** to address imbalanced discrete columns and improves stability and quality of generated data.

We used CTGAM model implemented in the [Modeling Tabular data using Conditional GAN](https://arxiv.org/abs/1907.00503) paper, presented at the 2019 NeurIPS conference.

The model utilized the mode-specific normalization, a conditional generator and training-by-sampling. It used fully-connected networks, applied Gumbel softmax and leaky ReLU activations, linear transformation, batch normalization and dropout. It used Wasserstein GAN loss with gradient penalty and Adam optimizer.

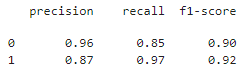
**Training Process:** The GAN training involves optimizing both networks to minimize the discriminator's ability to distinguish between real and synthetic data while maximizing the generator's ability to create convincing samples. Hyper parameters such as learning rate, batch size and number of training epochs were tuned for optimal performance.

### Deployment and Monitoring:

* + **MLOps Integration:** MLflow and Pyngrok were used to deploy the model and provide a user-friendly interface for experiment tracking. This setup facilitates continuous monitoring and rapid iteration of model versions as new data becomes available.
  + **Streamlit**: The final XGBoost model was deployed on streamlit to create intuitive, interactive web applications that enhance user engagement with models.
  + **Hugging Face and Gradio**: Hugging Face Spaces combined with Gradio simplifies model deployment, making it accessible for non-technical users.

**Results:**  
Key Findings:

* The integration of GANs significantly improved the model's ability to detect fraud by providing a richer and more diverse training dataset.



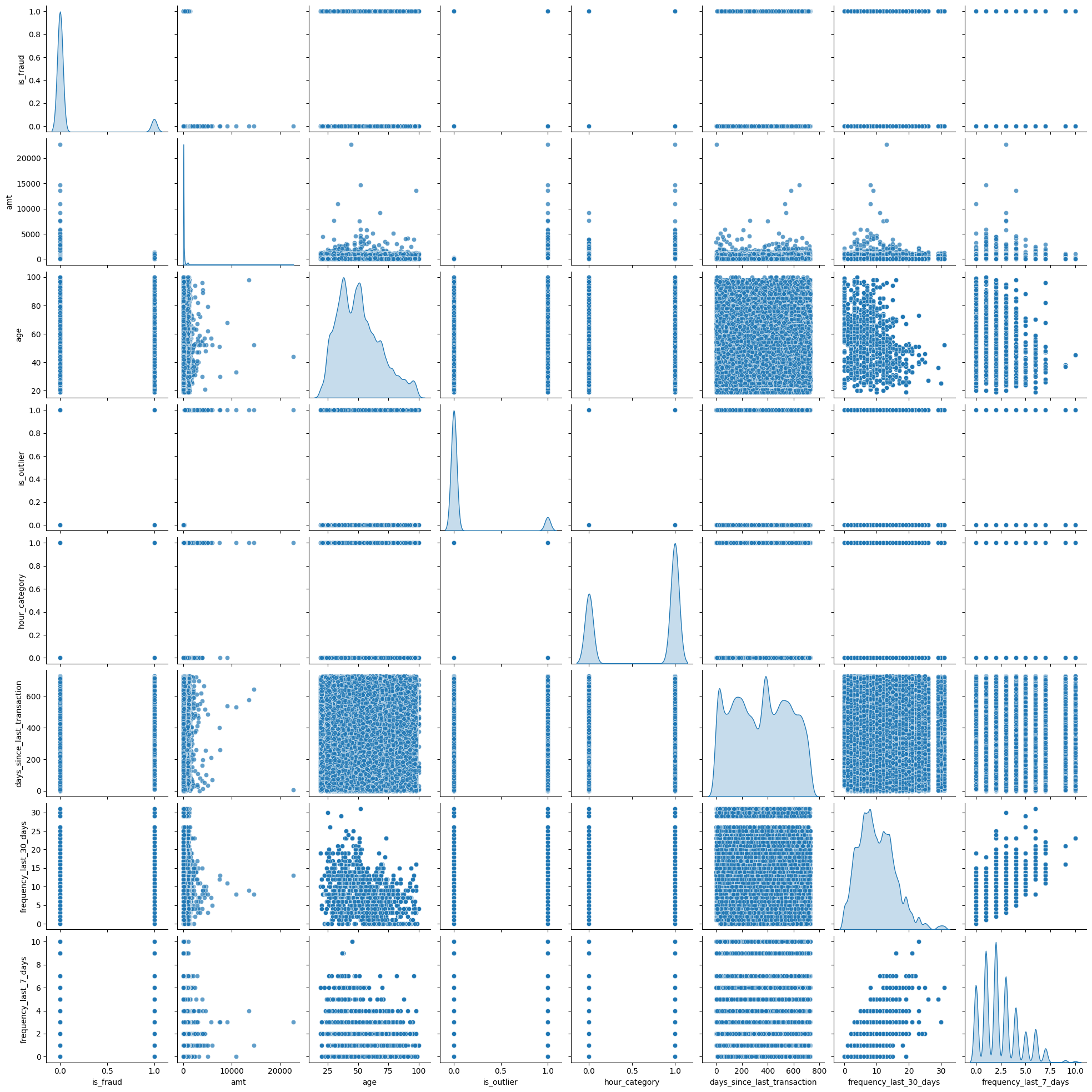
* The XGBoost model, leveraging engineered features, demonstrated a high recall for fraudulent transactions, indicating its effectiveness in minimizing false negatives, which are critical in fraud detection scenarios.

## Demonstration:

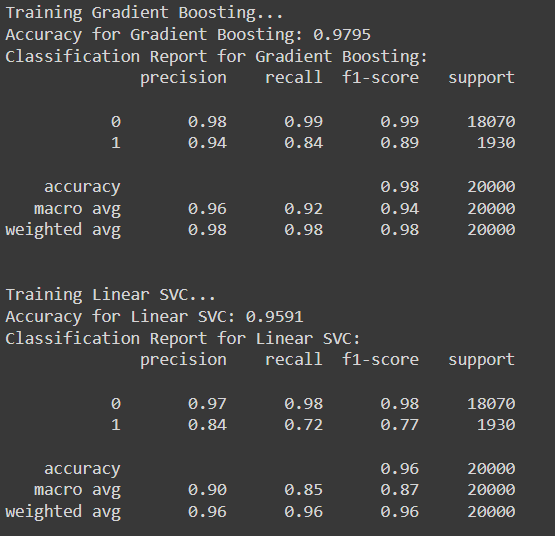
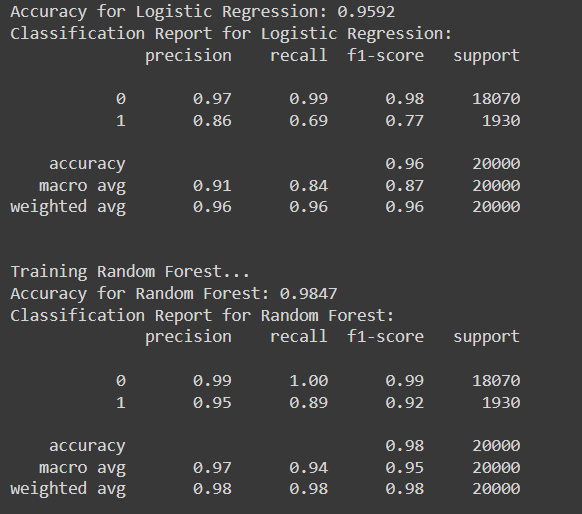
* A practical demonstration involved simulating a real-world environment where the model processed incoming transactions and flagged potential frauds. The system's responsiveness and accuracy were evaluated, confirming its suitability for deployment in financial institutions.

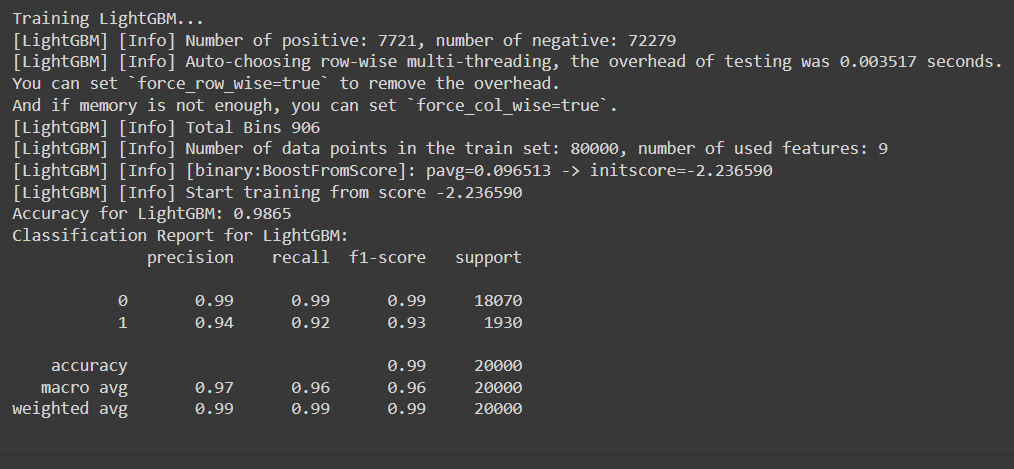
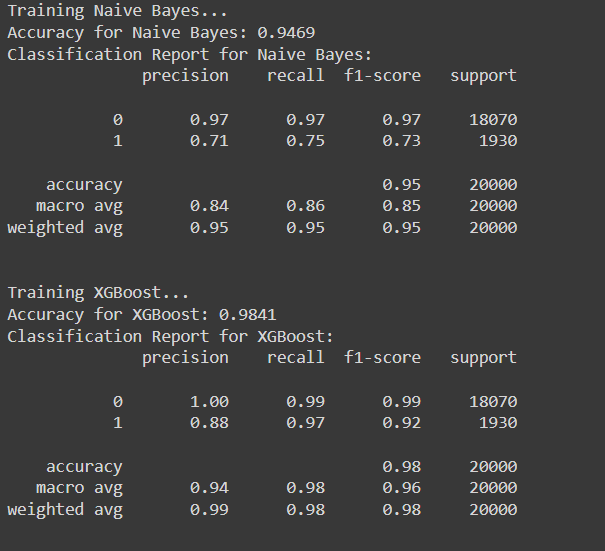
The following figures shows graphs of features importance and other graphs between features and each other, some features with the output column,





The following figure is the accuracy test performed on each model used:





**Challenges:**  
Difficulties Faced:

* Ensuring the quality of synthetic data generated by GANs was a significant challenge. The TCGAN needed fine-tuning regarding suitable number of training epochs to balance the discriminator and generator's learning pace effectively.
* Managing the volume of training data and computational demands of training GANs.

## How They Were Overcome:

* Iterative tuning of the GAN architecture, including layer configurations and learning rates, was conducted to produce high-quality synthetic data. Leveraging the innovative architecture of TCGAN from the literature helped approach the correct model.

**Future Work:**  
Improvements:

* Future enhancements include incorporating more sophisticated NLP techniques to analyze text-based transaction data, providing additional insights into potential fraud indicators.
* Exploring ensemble methods to combine the strengths of multiple classifiers may yield further improvements in detection accuracy.

## Next Steps:

* Deploying the model in a real-time fraud detection setup, where it continuously monitors transaction streams and adapts to new fraud patterns using online learning techniques.
* Expanding the dataset with additional features and sources, including social media and behavioral data, to improve the model's contextual understanding of transactions.

**Conclusion:**  
The developed fraud detection system effectively addresses the challenge of identifying fraudulent transactions in financial systems. By integrating advanced machine learning techniques and MLOps practices, the project delivers a robust solution capable of adapting to evolving fraud tactics. The use of GANs for data augmentation stands out as a novel approach, significantly enhancing the model's training process and overall performance, thus contributing to increased security and trust in digital financial transactions.