Credit Card Fraud Detection using Supervised Learning with Feature Engineering

K. T. Krishna Kumar[1], D. Satvika Reddy[2], K. Vasantha[3], P. Ramamohan[4], D. Veerendra Santosh Sai[5]

[1]*Associate Professor & Training and Placement officer, Computer Science And Engineering, Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India*

[2, 3, 4,54th] *B.Tech Final Semester, Bachelor of Technology, , Computer Science And Engineering, Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India*

---***---

Abstract - Credit card fraud causes significant financial losses for customers and financial institutions worldwide. Detecting fraudulent transactions accurately and at the earliest stage is therefore essential. This work presents a fraud detection framework based on supervised learning combined with effective feature engineering techniques. Historical transaction data containing both genuine and fraudulent operations are used to train predictive models. The objective is to classify new transactions with high reliability while minimizing false alarms. Credit card fraud detection is naturally formulated as a binary classification problem. The dataset is highly imbalanced, with fraudulent cases representing only a small fraction of the total records. To address this challenge, careful preprocessing and transformation of the data are performed. Feature engineering is applied to derive informative attributes that enhance the learning capability of the models. The proposed system evaluates multiple machine learning approaches within a unified pipeline. Model performance is assessed using standard evaluation metrics suited for imbalanced data. The framework aims to improve fraud capture rate without increasing unnecessary investigations. Experimental analysis demonstrates that engineered features significantly influence predictive accuracy. The study highlights the importance of combining domain knowledge with data-driven methods. The resulting solution provides a practical and scalable approach for real-world credit card fraud detection.

***Key Words*:** Credit Card Fraud Detection, Supervised Learning, Feature Engineering, Imbalanced Data, data preprocessing, Synthetic Minority Over-sampling Technique, Fraud Analytics, Model Evaluation

1.INTRODUCTION

Credit card usage has increased rapidly with the growth of digital commerce and online payment systems. [7] Along with this expansion, fraudulent activities have also grown in scale and sophistication. Financial institutions therefore require intelligent systems that can automatically distinguish legitimate transactions from malicious ones. Credit card fraud detection has become a critical research problem due to the severe financial losses and reputational risks involved. [20] Traditional rule-based monitoring systems are often inadequate because fraud patterns evolve continuously. As a result, machine learning approaches have gained significant attention for their ability to learn complex behaviors from historical data. In a real-world environment, fraud detection is challenging because fraudulent transactions represent only a very small percentage of the total volume. This class imbalance makes standard accuracy measures unreliable and demands more robust evaluation strategies. Moreover, customer spending behavior changes over time, introducing concept drift that can degrade model performance. Another difficulty is the limited availability of labeled fraud data, as investigation and confirmation require human effort. Hence, automated techniques must operate effectively even with scarce supervision. Supervised learning models provide a practical framework where past labeled transactions are used to predict the risk of new ones. [17]These models transform transaction attributes into meaningful features that help in identifying hidden patterns. Feature engineering and data preprocessing therefore play a vital role in improving predictive capability. Sampling strategies such as oversampling and undersampling are frequently employed to rebalance the dataset. By doing so, classifiers can better recognize minority fraud cases without being biased toward genuine transactions. The objective of this work is to design a data-driven fraud detection pipeline that supports investigators by prioritizing high-risk alerts. Multiple algorithms are evaluated to understand their strengths under highly imbalanced conditions [5]. Performance is analyzed using metrics that reflect real detection capability rather than simple overall correctness. The study also emphasizes the importance of continuously updating models with new feedback. Such adaptive systems are essential for maintaining effectiveness against emerging fraud strategies.

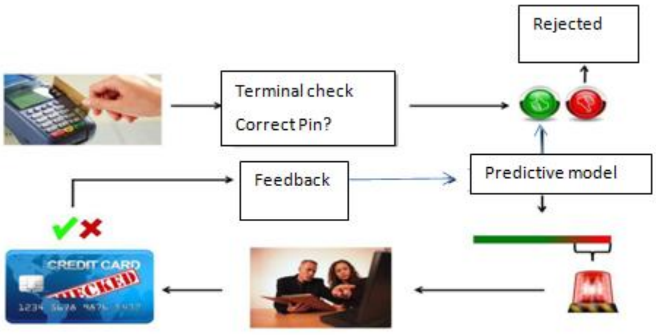


Fig.1. The Credit card Fraud Detection process

A customer initiates a card transaction and the system first checks basic details like PIN validity. The transaction information is then passed to a machine learning model, which evaluates whether it is genuine or fraudulent. Based on the prediction, the transaction is either approved or rejected, and the outcome is used as feedback to improve future detection.

2. Body of Paper

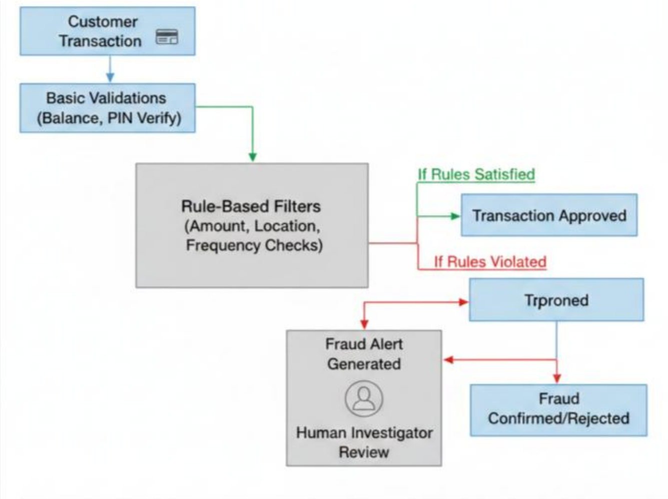
2.1 EXISTING SYSTEM

In traditional banking environments, fraud detection has largely relied on manual verification and predefined rule-based mechanisms. Transactions are typically filtered using static thresholds such as transaction amount, location, or frequency. [9] These rules are designed by domain experts based on previously observed fraud patterns. While simple to implement, rule-based approaches struggle to adapt to newly emerging strategies. Fraudsters continuously modify their behavior, making fixed rules less effective over time. As a result, many fraudulent transactions bypass detection while genuine ones may be incorrectly flagged.

Another limitation of existing systems is their heavy dependence on human investigators. Alerts generated by basic screening mechanisms require manual review, which increases operational cost and processing time. Since transaction volumes are extremely high, only a small subset can be examined in detail. [3] This constraint often leads to delayed responses and undetected fraud. Moreover, rule updates require expert intervention, which is both time-consuming and inflexible.

Some earlier solutions incorporated statistical techniques, but they often lacked scalability for large datasets. These methods typically used limited features and did not exploit the full richness of transaction information. Handling highly imbalanced data was also a major challenge, causing models to favor genuine transactions. Consequently, fraud detection rates remained low despite high overall accuracy. The absence of automated feature construction further restricted predictive capability.

In many legacy implementations, evaluation metrics focused primarily on accuracy rather than recall or precision. Such measurements can be misleading in fraud scenarios where minority detection is critical. Furthermore, continuous learning from investigator feedback was not effectively integrated. Without adaptation, system performance deteriorates as patterns change.



CHALLENGES

- Fraud cases are very rare compared to normal transactions.
- Getting confirmed fraud labels is difficult and expensive.
- Fraudsters keep changing methods, reducing model accuracy over time.
- Genuine customers may be wrongly blocked.
- Finding useful indicators of fraud is not easy.
- Missing or noisy data can mislead the model.
- Decisions must be made within seconds.
- Systems must handle huge volumes of daily transactions.
- Investigators need clear reasons for predictions.
- Different errors lead to different financial impacts.

2.2 PROPOSED SYSTEM

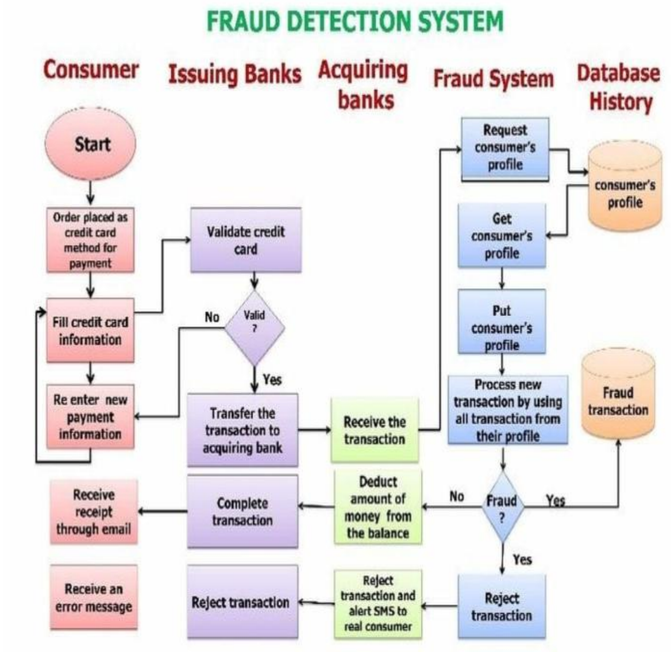
The proposed system presents a data-driven framework for identifying fraudulent credit card transactions using supervised machine learning techniques. The objective is to automatically distinguish illegitimate activities from genuine customer behavior with high reliability. [15] Historical transaction records containing labeled outcomes are used as the foundation for model training. Each incoming transaction is treated as a data instance that must be classified in real time. Preprocessing operations are applied to clean the dataset and prepare it for effective learning.

The system performs normalization, missing value handling, and noise reduction to improve data quality. Feature engineering techniques are employed to convert raw transaction attributes into meaningful predictive variables. These derived features capture behavioral patterns, spending habits, and risk indicators. [2] By enhancing representation, the model becomes more capable of separating fraud from legitimate activity. Class imbalance handling methods are also integrated to ensure minority fraud cases receive adequate attention.

Multiple supervised learning algorithms are trained and validated on historical data. The classifiers learn decision boundaries that help predict whether a new transaction is suspicious. [11]During deployment, the trained model evaluates unseen transactions and generates risk scores. Based on these scores, alerts can be triggered for further investigation. This automated process significantly reduces manual workload.

Performance evaluation is conducted using appropriate metrics such as precision, recall, F1-score. [1] These measures provide a realistic understanding of detection capability under imbalanced conditions. Continuous monitoring enables periodic retraining as new fraud strategies emerge. The system architecture is designed for scalability to handle large transaction volumes efficiently.

Experimental results demonstrate that the proposed approach improves fraud detection accuracy while minimizing false alarms. The framework supports adaptability, transparency, and integration with existing banking infrastructure. [6] Overall, the solution provides an intelligent and practical mechanism for strengthening financial security and customer trust.


FRAUD DETECTION SYSTEM
Consumer
Issuing Banks
Acquiring banks
Fraud System
Database History
Start
Order placed as credit card method for payment
Validate credit card
Request consumer's profile
consumer's profile
Get consumer's profile
Put consumer's profile
Fill credit card information
No
Valid ?
Yes
Re enter new payment information
Transfer the transaction to acquiring bank
Receive the transaction
Process new transaction by using all transaction from their profile
Fraud transaction
Receive receipt through email
Complete transaction
Deduct amount of money from the balance
No
Fraud ?
Yes
Yes
Receive an error message
Reject transaction
Reject transaction and alert SMS to real consumer
Reject transaction

ADVANTAGES

- Provides higher fraud detection accuracy by learning patterns from historical labeled data.
- Reduces manual effort by automatically generating risk scores for new transactions.
- Handles imbalanced datasets effectively through sampling and feature engineering methods.
- Adapts to new fraud strategies via periodic retraining and model updates.
- Minimizes financial loss and customer inconvenience by lowering false alarms.

2.3 Algorithms and Techniques

To identify an effective model for credit card fraud detection, multiple supervised machine learning algorithms are evaluated under the same experimental setup[18]. All models are trained and tested on the same preprocessed and feature-engineered dataset to ensure a fair comparison. Performance is assessed using standard evaluation metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. This comparative study helps determine the most reliable and scalable approach for identifying fraudulent transactions in real-world environments [13]. In this project, Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, and XGBoost are implemented and analyzed.

The following algorithms and techniques are used in this project for fraud detection and transaction classification:

- Logistic Regression.
- K-Nearest Neighbors.
- Random Forest.
- Decision Tree.
- XGBoost.

A. Logistic Regression

Logistic Regression is a supervised learning algorithm widely used for binary classification tasks. It predicts the probability of a transaction being fraudulent or genuine using a logistic (sigmoid) function. [10] The model is computationally efficient and easy to interpret. It serves as a strong baseline for comparison with more complex approaches. Logistic Regression performs well when relationships between features and output are relatively linear.

B. K-Nearest Neighbors

K-Nearest Neighbors classifies a transaction based on similarity to previously observed data points. The class is assigned according to the majority vote among the nearest neighbors. [14] It is a simple and intuitive method that can capture nonlinear boundaries. Proper feature scaling is important for achieving good results. The effectiveness of the model depends on the choice of K and the distance metric.

C. Decision Tree

Decision Tree is a rule-based supervised learning technique that splits the dataset according to feature conditions. Each branch represents a decision rule, and leaf nodes correspond to class labels. The model is easy to visualize and interpret, making it suitable for explaining fraud predictions. It can automatically select important features. However, individual trees may suffer from overfitting.

D. Random Forest

Random Forest is an ensemble learning approach that builds multiple decision trees and combines their outputs. It introduces randomness in both data sampling and feature selection. [4] This helps improve prediction accuracy and control overfitting. The method is robust to noise and works well on imbalanced datasets. It is widely adopted in financial fraud detection applications.

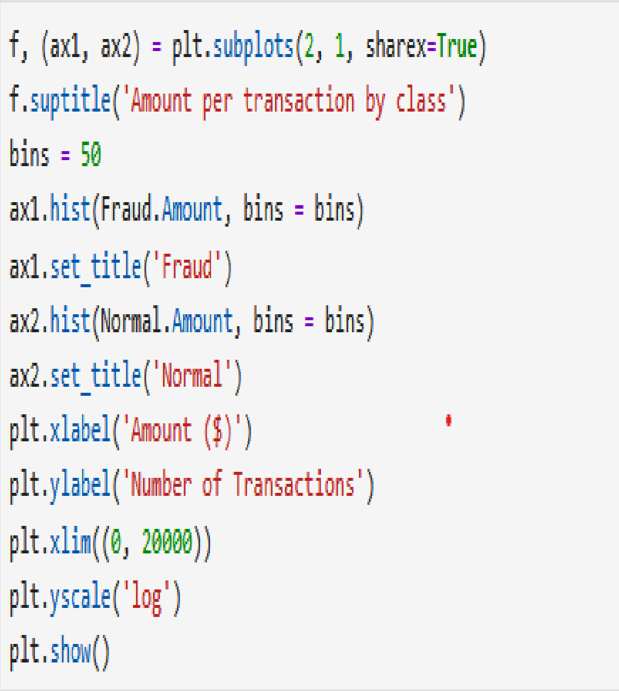
E. XGBoost

XGBoost is a powerful gradient boosting technique known for its speed and high predictive performance. It builds models sequentially, where each new model focuses on correcting previous errors. Regularization mechanisms enhance generalization capability. [12] XGBoost efficiently handles complex patterns in transaction data. It often achieves superior results compared to traditional classifiers.

2.4 METHODOLOGY

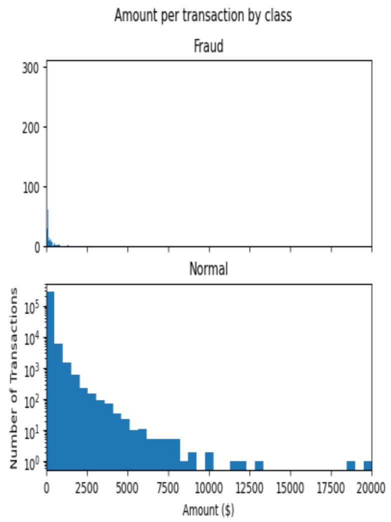
INPUT:

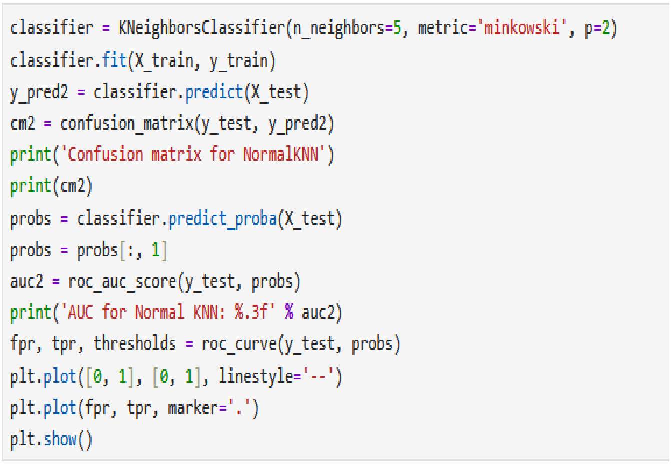
The input data for this credit card fraud detection system is sourced from the European credit card transaction dataset (September 2013), [19] comprising 284,807 transactions. Due to confidentiality, 28 features (V1–V28) are provided as PCA-transformed principal components, with only 'Time' (seconds elapsed since first transaction) and 'Amount' (transaction value in euros) remaining untransformed . The binary response variable 'Class' indicates fraud (1) or genuine (0), with frauds representing only 0.172% of transactions (492 cases), exhibiting severe class imbalance. [8] Preprocessing includes normalization of the 'Amount' feature and scaling of PCA components. No missing values are present. These processed inputs enable machine learning models to identify fraudulent transaction patterns.



```
f, (ax1, ax2) = plt.subplots(2, 1, sharex=True)
f.suptitle('Amount per transaction by class')
bins = 50
ax1.hist(Fraud.Amount, bins = bins)
ax1.set_title('Fraud')
ax2.hist(Normal.Amount, bins = bins)
ax2.set_title('Normal')
plt.xlabel('Amount ($)')
plt.ylabel('Number of Transactions')
plt.xlim((0, 20000))
plt.yscale('log')
plt.show()
```

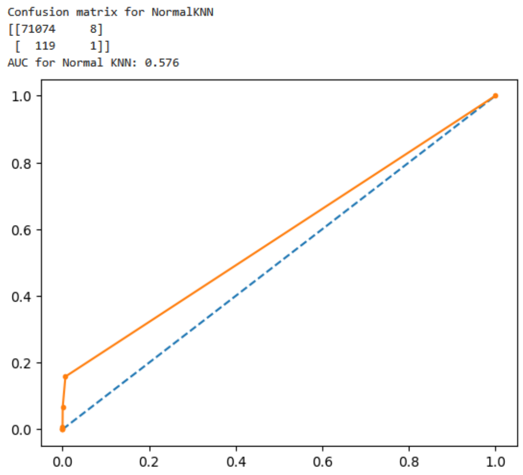
Output:





```
classifier = KNeighborsClassifier(n_neighbors=5, metric='minkowski', p=2)
classifier.fit(X_train, y_train)
y_pred2 = classifier.predict(X_test)
cm2 = confusion_matrix(y_test, y_pred2)
print('Confusion matrix for NormalKNN')
print(cm2)
probs = classifier.predict_proba(X_test)
probs = probs[:, 1]
auc2 = roc_auc_score(y_test, probs)
print('AUC for Normal KNN: %.3f' % auc2)
fpr, tpr, thresholds = roc_curve(y_test, probs)
plt.plot([0, 1], [0, 1], linestyle='--')
plt.plot(fpr, tpr, marker='.')
plt.show()
```

Output:



```
Confusion matrix for NormalKNN
[[71074     8]
 [  119     1]]
AUC for Normal KNN: 0.576
```

3. CONCLUSIONS

This project evaluated five machine learning algorithms for credit card fraud detection on highly imbalanced data (0.172% fraud). XGBoost outperformed all classifiers with the highest AUC-ROC of 0.883. SMOTE and SMOTE+Tomek links proved most effective for handling class imbalance, while cluster centroids performed worst. Results demonstrate that frequent model updates (weekly/biweekly) significantly improve detection accuracy over static annual updates. The optimal configuration combines XGBoost with SMOTE-based oversampling, providing a robust framework for real-world fraud detection systems.

ACKNOWLEDGEMENT

The authors express their sincere gratitude to Sunil Kumar, Principal of Sanketika Vidya Parishad Engineering College, for providing the opportunity to carry out this project. They thank T. Sunil Kumar, Head of the Department of CSE, for the continuous support and encouragement. The authors are especially indebted to their project guide, K. T. Krishna Kumar, for his valuable guidance and supervision throughout the work. They also acknowledge the faculty members of the Department of CSE for their assistance. Finally, the authors thank their family and friends for their constant motivation and support

REFERENCES

[1] Avinash Dua & Vishal A. Gahlaut, “Credit Card Fraud Destection Using Machine Learning” – International Journal of Modern Developments in Engineering and Science (2025). Supervised ML models with feature engineering handling class imbalance. https://journal.ijmdes.com/ijmdes/article/view/268

[2] Emmanuel Ileberi, Yanxia Sun & Zenghui Wang, “A machine learning based credit card fraud detection using GA for feature selection” – Journal of Big Data (2022). Genetic algorithm for feature selection and multiple supervised classifiers. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-022-00573-8

[3] Viraj Soni & Sumit Gupta, “Fraud Detection in Credit Card Transactions: A Machine Learning Approach” – Journal of Electrical Systems (2024). ML models with emphasis on feature preprocessing &evaluation. https://journal.esrgroups.org/jes/article/view/8318

[4] Varun Kumar K S et al., "Credit Card Fraud Detection Using Machine Learning Algorithms" – IJERT (2020). Compares traditional supervised classifiers on fraud data. https://www.ijert.org/credit-card-fraud-detection-using-machine-learning-algorithms

[5] Hadiza Hassan et al., "An Enhanced Feature Engineering Technique for Credit Card Fraud Detection" – FUDMA Journal of Sciences (2024). Hybrid sampling & enhanced engineering for better classification. https://fjs.fudutsinma.edu.ng/index.php/fjs/article/view/2435

[6] P. K. R. et al., "A Comparative Analysis of Machine Learning Techniques for Detecting Credit Card Fraud" – IJISAE (2024). Model comparison with engineered features.
https://ijisae.org/index.php/IJISAE/article/view/5232

[7] S. Bhuvaneswar et al., "A Supervised ML Algorithm for Detecting and Predicting Fraud Credit Card Transactions" – IRJAEH (2024). Framework for supervised detection. https://irjaeh.com/index.php/journal/article/download/409/373/830

[8] A. Correa Bahnsen et al., "Feature Engineering Strategies for Credit Card Fraud Detection" – Expert Systems with Applications (2016). Classic strategies for feature aggregation and behavior features.

[9] Xuetong Niu, Li Wang & Xulei Yang, "A Comparison Study of Credit Card Fraud Detection: Supervised vs Unsupervised" – arXiv (2019). Supervised vs unsupervised performance evaluation.
https://arxiv.org/abs/1904.10604

[10] Yvan Lucas et al., "Multiple Perspectives HMM-based Feature Engineering for Credit Card Fraud Detection" – arXiv (2019). Automated feature engineering for sequence modeling.
https://arxiv.org/abs/1905.06247

[11] A Patel, "Exploring Supervised Machine Learning Techniques… for Credit Card Fraud Detection" – ITM Conf. (2024). Highlights imbalance challenges.

[12] M. S. Prateeksha et al., "Fraud Detection Using Supervised Learning Algorithms" – ResearchGate article (2023). Compares multiple classifiers.

[13] KG Dastidar, "Machine Learning Methods for Credit Card Fraud Detection" – Opus4 survey (2024). ML overview including preprocessing & feature selection.

[14] Ibrahim Y. Hafez et al., "A Systematic Review of AI-Enhanced Techniques in Credit Card Fraud Detection" – Journal of Big Data (2025). Reviews supervised and deep learning approaches. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-01048-8

[15] Zahra Faraji, "A Review of Machine Learning Applications for Credit Card Fraud Detection with A Case Study" – SEISENSE Journal of Management (2022). Supervised models & metrics review.

[16] Reethu Elza Joseph & Dr. L. C. Manikandan, "A Review on Credit Card Fraud Detection Techniques" – IJERT (2022). Broad survey article.

[17] Harsh Mehta, "Analysis of Different Machine Learning Models for Credit Card Fraud Detection" – IRJET (2024). Comparative performance insights + feature use.

[18] "Credit Card Fraud Detection Using Feature Fusion-based ML Model", Qing Meng – HSET (3-model fusion & feature processing).

[19] "A Novel Credit Card Fraud Detection Feature Selection System" – ACM (2024). Innovative preprocessing and feature selection.

[20] Feature Engineering Strategies (PDF) – Bahnsen's expanded feature set including behavioral & periodic features.

[21] Cost-Sensitive Machine Learning Theory – Wikipedia reference useful for handling class imbalance in supervised systems.

[22] Feature Learning & Supervised Representations – Wikipedia overview (context for automating feature construction).

[23] "Credit Card Fraud Detection Using Machine Learning and Data Science" – S.P. Maniraj et al., IJERT (2019) – Early ML fraud detection baseline.

[24] "Credit Card Fraud Detection using ML Algorithms: IJERT Paper" – foundational 2020 article for classifier comparisons.

[25] "Feature Engineering in Fraud Detection" – Medium article summarizing key techniques (useful for background).

BIOGRAPHIES

K. Tulasi Krishna Kumar is a Ratified Associate Professor from Andhra University and the Placement Officer of SVPEC, with 16 years of extensive experience in training and placing students in IT, ITES, and core industry sectors. He has successfully trained more than 16,000 students and 750 faculty members through FDP & sessions. He has authored 8 books and has guided over 85 UG and PG project teams. A Certified Campus Recruitment Trainer (JNTUA), he holds an M.Tech degree in Computer Science and Engineering and is currently pursuing his Ph.D. in CSE.

D. Satvika Reddy is a final-year B.Tech student at Sanketika Vidya Parishad Engineering College, accredited with an A grade by NAAC and affiliated with Andhra University. Her interests include machine learning and fraud analytics for secure financial systems. She is working on her major project, CREDIT CARD FRAUD DETECTION USING SUPERVISED LEARNING WITH FEATURE ENGINEERING, focusing on class imbalance handling and performance improvement through model comparison. The project is carried out under the guidance of Kandhati Tulasi Krishna Kumar Nainar, Assistant Professor, Training & Placement Officer, SVPEC.

K. Vasantha is a final-year B.Tech student at Sanketika Vidya Parishad Engineering College, accredited with an A grade by NAAC and affiliated with Andhra University. Her interests include machine learning and fraud analytics for secure financial systems. She is working on her major project, CREDIT CARD FRAUD TECTION USING SUPERVISED LEARNING WITH FEATURE ENGINEERING, focusing on class imbalance handling and performance improvement through model comparison. The project is carried out under the guidance of Kandhati Tulasi Krishna Kumar Nainar, Assistant Professor, Training & Placement Officer, SVPEC.

D. Veerendra Santosh Sai is a final-year B.Tech student at Sanketika Vidya Parishad Engineering College, accredited with an A grade by NAAC and affiliated with Andhra University. His interests include machine learning and fraud analytics for secure financial systems. He is working on his major project, CREDIT CARD FRAUD DETECTION USING SUPERVISED LEARNING WITH FEATURE ENGINEERING, focusing on class imbalance handling and performance improvement through model comparison. The project is carried out under the guidance of Kandhati Tulasi Krishna Kumar, Assistant Professor, Training & Placement Officer, SVPEC.

P. Rama Mohan is a final-year B.Tech student at Sanketika Vidya Parishad Engineering College, accredited with an A grade by NAAC and affiliated with Andhra University. His interests include machine learning and fraud analytics for secure financial systems. He is working on his major project, CREDIT CARD FRAUD DETECTION USING SUPERVISED LEARNING WITH FEATURE ENGINEERING, focusing on class imbalance handling and performance improvement through model comparison. The project is carried out under the guidance of Kandhati Tulasi Krishna Kumar, Assistant Professor, Training & Placement Officer, SVPEC.