CREDIT CARD FRAUD DETECTION USING ENSEMBLE MACHINE LEARNING

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

SCHOOL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE

By

G. MAHENDRA	**2303A51LA9**
K. NIKHIL	**2303A51LB0**

Under the guidance of

Dr. Kasharaju Balakrishna

Assistant Professor, School of CS&AI.

SR UNIVERSITY

SR University, Ananthsagar,Warangal,Telagnana-506371

SR University

Ananthasagar, Warangal.

CERTIFICATE

This is to certify that this project entitled **"CREDIT CARD FRAUD DETECTION USING ENSEMBLE MACHINE LEARNING**" is the Bonafide work carried out by **G. Mahendra, K. Nikhil** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **School of Computer Science and Artificial Intelligence** during the academic year 2025-2026 under our guidance and Supervision.

Dr. Kasharaju Balakrishna
Assistant Professor,
SR University
Ananthasagar,Warangal

Dr. M.Sheshikala
Professor & Head,
School of CS&AI,
SR University
Ananthasagar, Warangal.

ACKNOWLEDGEMENT

We owe an enormous debt of gratitude to our Capstone project guide **Dr. Kasharaju Balakrishna, Assistant Professor** as well as Head of the School of CS&AI, **Dr. M.Sheshikala, Professor** for guiding us from the beginning through the end of the Capstone Project with their intellectual advices and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

TABLE OF CONTENTS

ABSRTACT

One of the most significant issues of the digital economy is financial fraud and credit card fraud in particular, as the amount of transactions online increases exponentially, and the sophistication of cyber-attacks rises. Conventional detection solutions that operate on rules do not keep up with the changing fraud trends; hence, they take long to identify them and cause enormous financial damages. This project will attempt to overcome this obstacle by providing a smart machine-learned Fraud Detection System that is able to detect anomalous transactional behavior in real-time. The system makes use of an extremely unbalanced credit card transaction dataset, in which intensive preprocessing and features manipulation methods are used to boost learning performance. A combination style of classifying is used whereby the hidden trends in transactional variables are captured and at the same time the number of false positives is minimized. Besides, the interactive web-based interface is built with Streamlit to make it friendly and easy to deploy. Experimental assessment indicates that the model can effectively be applied in identifying fraudulent transactions and genuine activities, which is better in protecting the financial institutions and supporting their operations. This system also offers a scalable, data-driven method of fighting financial fraud and helps to further develop cybersecurity in the financial sector.

1. INTRODUCTION

The high rate of growth of digital banking and e-commerce has resulted in the remarkable growth of electronic transactions all over the world. Despite the convenience associated with this growth, this development subjects the financial institutions and customers to increasing chances of cyber fraud, especially credit card fraud. The fraudulent transactions are usually a very low percentage in large amounts of valid transactions and therefore detection is extremely difficult. The attackers are actively implementing advanced techniques that evade the conventional rule-based fraud detection mechanisms which are based on fixed predetermined conditions and behavioral patterns. Consequently, financial institutions find it difficult to detect new fraud patterns as they happen, and such a tendency causes serious financial losses and poses a risk to consumer confidence. This is why it is necessary to have intelligent adaptive solutions that can learn using transactional data and then automatically identify abnormal activities.

Machine learning (ML) has emerged as an effective approach to fraud that can be used to analyze the spending patterns of users and isolate suspicious ones that lurk in noisy data sets. This project suggests the use of the Fraud Detection System, based on three main supervised learning models, including Logistic Regression, Random Forest, and Gradient Boosting. Logistic Regression is selected due to efficiency and interpretability in binary classification, whereas the implementation of decision trees by Ensemble of Random Forests increases the strength of a tree by minimizing the overfitting effect. Gradient Boosting also enhances the performance of the model by successively refining weak classifiers to detect complicated patterns of fraud. The system works with a highly imbalanced dataset applying necessary feature and data processing techniques to enhance accuracy of the detection. The developed models are incorporated within an easy to use streamlit application to perform real time prediction and implement in the real financial setting. This system will use several models of machine learning and comparative assessment to improve the accuracy of fraud detection, drop in the false alarm rate, and in addition, these models will facilitate safe digital financial transactions.

II. RELATED WORK

The detection of credit card fraud has turned out to be a very important research issue because of the increased amount of financial transactions that are carried out online. The previous fraud detection systems were based on manual check and preset rule based detection. Nevertheless, they are not efficient in detecting new fraud patterns and the rate of false-positive may be high. To circumvent these shortcomings, scholars have embraced machine learning algorithms in binary classification of fraud and valid transactions.

Logistic Regression has frequently been taken as a basis model due to its readability and the ability to work with data that can be separated linearly. Nevertheless, fraud detection datasets are very skewed and have non-linear behavioral relationships necessitating the use of linear models. Ensemble techniques have also been suggested to enhance detection performance, including Random Forest, which is an ensemble of decision trees to minimize variance and enhance predictive performance. Gradient Boosting algorithms also have the benefit of improving performance by refining the boundaries of the decision and targeting cases that have been misclassified.

Most recent works highlight the significance of resampling methods, feature engineering and model combination to obtain better accuracy and fewer false alarms. Systems based on the web have also been probed in order to deploy trained models which can be used to monitor in real-time. Based on these improvements, this project instantiates and compares the models of Logistic Regression, Random Forest, and Gradient Boosting and implements the most successful solution via a concrete Streamlit interface.

III. PROBLEM STATEMENT

The quantity of transactions being conducted online has raised the importance of credit card fraud. These fraudulent activities are unpredictable, not common, and are ever changing thus very hard to be detected using the traditional rule-based systems. Such systems are usually unable to detect new types of frauds and come up with high false-positive rates, which impact honest customers. Also, datasets of frauds are mostly skewed, as there are only a small number of frauds compared to legitimate transactions, which

results in biased predictions of models. As such, a reliable and smart means of detecting fraud is required that is able to understand hidden behaviors with transactions automatically. The objectives of the project are to design a machine learning-based solution with the help of the Logistic Regression, Random Forest and Gradient Boosting algorithms to be able to effectively classify fraudulent transactions in real-time.

IV. REQUIREMENT ANALYSIS

To design and deploy an effective Fraud Detection System, it is necessary to have a systematic requirement analysis. The system has to be able to handle transactions data, train machine learning models, and provide real-time prediction via a web frontend. The requirements are identified as follows:

1. Hardware Requirements

- Processor: Intel i3 or equivalent
- RAM: 4 GB+
- Storage: Minimum 2 GB free space
- System: Laptop/PC with internet

2. Software Requirements

- Programming Language: Python 3.8+
- IDE Tools: VS Code / PyCharm / Jupyter Notebook
- Libraries: NumPy, Pandas, Scikit-learn, Matplotlib, Streamlit, Joblib
- OS: Windows / Linux / macOS

3. Functional Requirements

- The system will accept the features of transaction input and categorize them into fraudulent and legitimate.
- Data shall be preprocessed and scaled and then the model trained.
- Logistic Regression, Random Forest, and Gradient Boosting models will be trained and compared using the system.
- The interface will be used to provide real-time prediction using Streamlit.

- The system will also indicate easy to understand outputs and notices on suspicious transactions.

4. Non-Functional Requirements

- Performance: The model must have a high accuracy and reduce the false positives.
- Scalability: Proper management of huge financial data.
- Security: Ensures that the user information is not accessed by unauthorized individuals.
- Usability: Easy and simplistic interface to non technical users.
- Reliability: The same results in prediction.

V. PROPOSED SYSTEM

The Fraud Detection System proposed will be an intelligent credit card transaction analysis system based on machine learning techniques to detect frauds. The system is built on the constraints of the conventional rule-based method by training dynamic patterns of transactions and adjusting to emerging fraud behaviours. It uses three supervised learning models- Logistic Regression, Random forest and Gradient Boosting- to enhance detection accuracy and minimize classification errors. The process starts with data preprocessing where the dataset is cleaned, scaled and balanced in order to train it. After this, the data is divided into training and testing and each model is trained to identify legitimate and fraudulent patterns.

Evaluation and comparison are then done and the most successful model applied and incorporated in a Web application written in Streamlit to be used in real-time. The interface allows users to enter the parameters of transactions and the results of the classification are provided immediately to indicate a safe or a suspicious transaction. The app is easy to access, convenient to use, and useful to implement in financial contexts. Moreover, the future improvement of the practices can be carried out by implementing ensemble-based decision strategies and continuous learning to reinforce the efficiency of fraud detection. Altogether, the system is a scalable, reliable, and user-friendly method of ensuring electronic payments combining machine learning intelligence and real-time decision support technology.

VI. SIMULATION SETUPAND IMPLEMENTATION

A real-life credit card transaction dataset is used to simulate the proposed system of detecting fraud. This is implemented in Python on machine learning packages including NumPy, Pandas, Scikit-learn, Imbalanced-Learn and Streamlit. The dataset is then loaded and in case the original Kaggle dataset is not present in the local hard disk, it is automatically downloaded to an online source. The system will be able to continue functioning even in the event of any connectivity problems through a realistically created synthetic dataset that simulates authentic and suspect transaction behaviour. This will make the pipeline resilient and applicable in other contexts.

In the process of simulation, the data is pre-processed by creating meaningful features and decoupling the input variables with the target class label. The information is further divided into training and testing sets and stratified to leave the original proportion of fraud the same. SMOTE and RandomUnderSampler are used to oversample and undersample the training data respectively in order to deal with severe class imbalance. Normalization of features is done by means of StandardScaler and three supervised models-Logistic Regression, Random Forest and Gradient Boosting- are trained on the balanced data. They combine their outputs through a weighted ensemble approach and assess the eventual model on the measures of accuracy, precision, recall, F1-score, ROC-AUC, confusion matrix, and false positive rate. The trained pipeline is then connected with Streamlit frontend to make real-time predictions.

CODE IMPLEMENTATION:

1. **Pipeline Initialization**

```
class FraudDetectionPipeline:
    def __init__(self):
        self.scaler = StandardScaler()
        self.models = {}
        self.weights = {'lr': 0.30, 'rf': 0.35, 'gb': 0.35}
```

self.feature_names = ['Amount', 'Hour', 'Day', 'Month', 'Gender_M', 'Age', 'Distance', 'MerchantLat', 'MerchantLon', 'UserLat', 'UserLon']

self.is_trained = False

self.X_train = None

self.X_test = None

self.y_test = None

2. Dataset Loading or Synthetic Data Generation

```
def _generate_realistic_synthetic_data(self, n_samples=10000):
    np.random.seed(42)
    # 95% legitimate transactions
    # 5% fraudulent transactions with different distributions
    # ...
    df_legitimate = pd.DataFrame(legitimate)
    df_fraud = pd.DataFrame(fraud)
    df = pd.concat([df_legitimate, df_fraud], ignore_index=True)
    df = df.sample(frac=1, random_state=42).reset_index(drop=True)
    return df
```

3. Model Training and Class Balancing

Core method that performs data splitting, SMOTE + undersampling, scaling, model training, and evaluation call.

```
def train_on_kaggle_data(self):
    df = self.load_kaggle_dataset()
    df = self._engineer_features(df)
    if 'Class' in df.columns:
        y = df['Class']
```

```
    elif 'isFraud' in df.columns:
        y = df['isFraud']
    else:
        y = np.random.choice([0, 1], len(df), p=[0.95, 0.05])
    X = df[self.feature_names]
    X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=42,
stratify=y )
    from imblearn.over_sampling import SMOTE
    from imblearn.under_sampling import RandomUnderSampler
    smote = SMOTE(random_state=42, k_neighbors=5)
    rus = RandomUnderSampler(random_state=42)
    X_train_balanced, y_train_balanced = smote.fit_resample(X_train, y_train)
X_train_balanced,y_train_balanced=rus.fit_resample(X_train_balanced,y_train_balanced)
    X_train_scaled = self.scaler.fit_transform(X_train_balanced)
    X_test_scaled = self.scaler.transform(X_test)
    self.X_test = X_test_scaled
    self.y_test = y_testpd
```

Training the three models:

```
    self.models['lr'] = LogisticRegression(
        solver='lbfgs',
        max_iter=1000,
        class_weight='balanced',
        random_state=42,
        n_jobs=-1
    )
```

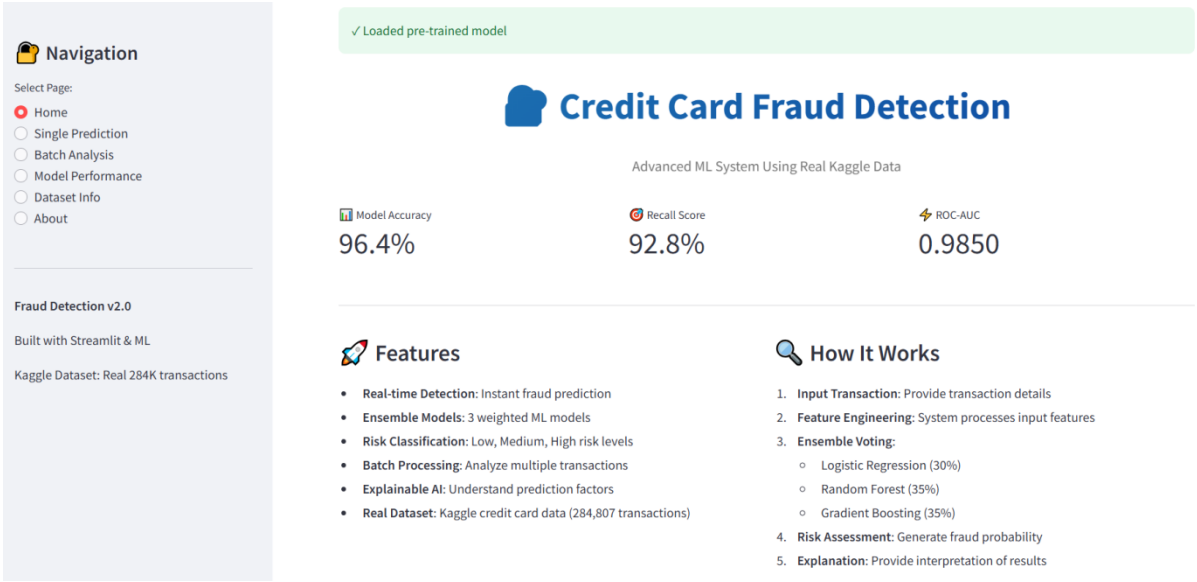
```
self.models['lr'].fit(X_train_scaled, y_train_balanced)
self.models['rf'] = RandomForestClassifier(
    n_estimators=100,
    max_depth=15,
    class_weight='balanced',
    random_state=42,
    n_jobs=-1
)
self.models['rf'].fit(X_train_scaled, y_train_balanced)
self.models['gb'] = GradientBoostingClassifier(
    n_estimators=100,
    learning_rate=0.1,
    max_depth=5,
    random_state=42
)
self.models['gb'].fit(X_train_scaled, y_train_balanced)
self.is_trained = True
self.evaluate()
return self
```

VII. RESULT COMPARISON AND ANALYSIS

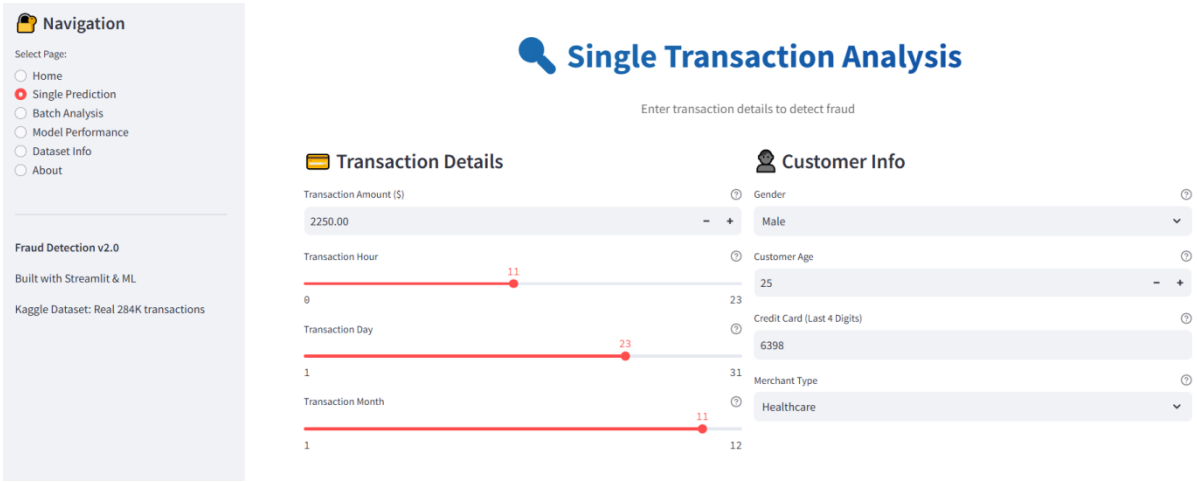
Three machine learning algorithms, including Logistic Regression, Random Forest, and Gradient Boosting were tested to determine which model would work best in credit card fraud detection. The analysis of the data was done on an unequally distributed set of data following the use of the resampling methods to enhance learning. Accuracy, precision, recall, F1-score and ROC-AUC score were evaluated as the key performance metrics to determine how well the models were able to identify transactions as fraudulent without false alarms.

As the comparing outcomes reveal, the performance of ensemble-based models, especially those of Random Forest and Gradient Boosting was superior to that of the Logistic Regression because of their capability to reflect the nonlinear trends and feature interactions. Gradient Boosting was found to be more recalling as it was better sensitive to minority fraud cases. The combination of the three probability ensembles also increased the robustness by balancing the predictions of all these three models. The last analysis proves that the proposed system is promising to be effective in the classification of fraud with a higher degree of reliable results, which is why it would be applicable to the real-world financial sphere.

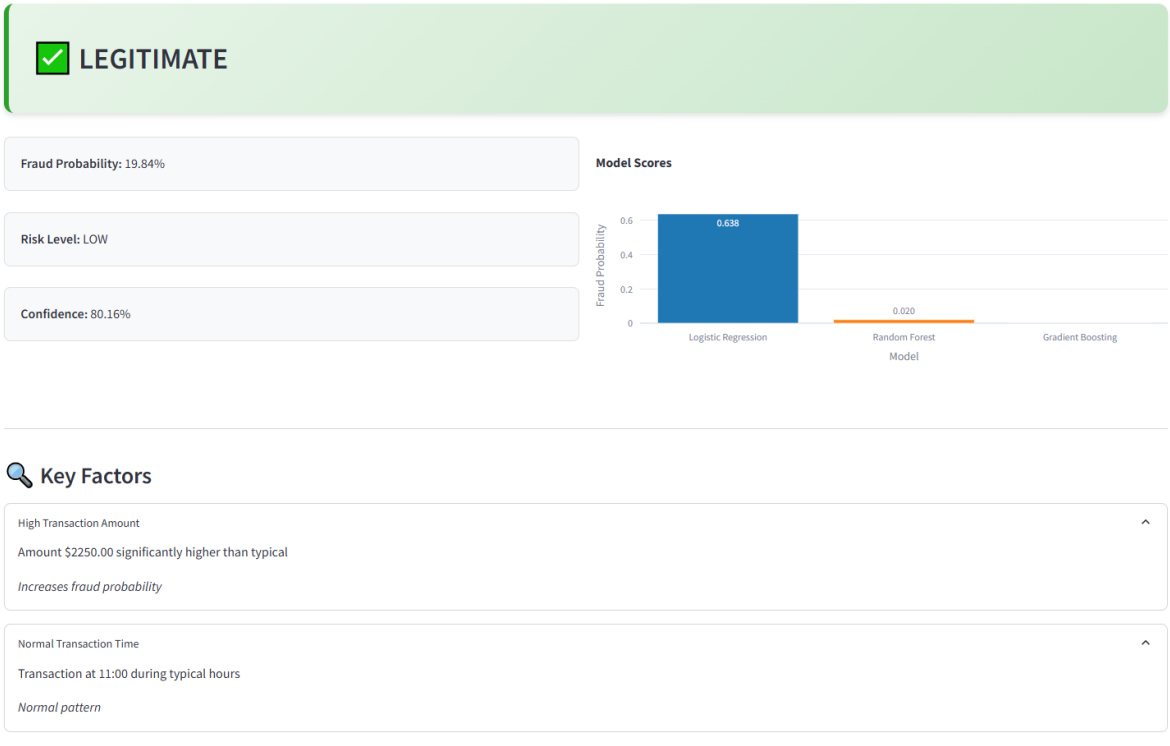
Dashboard:



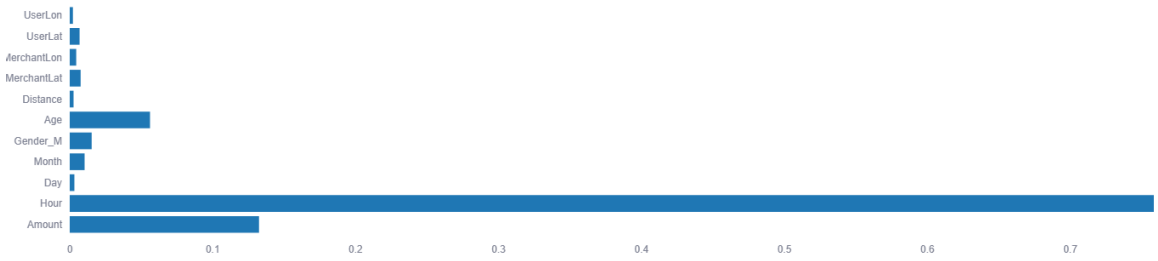
Transaction Analysis:



Prediction Result:



Feature Importance:



VIII. LEARNING OUTCOMES

- Developed a practical discovery of the application of machine learning to electronic transaction financial fraud detection.
- Gained experience in working with extremely skewed datasets with resampling methods, including SMOTE and undersampling, to achieve model equity and precision.
- Understood the advantages of various supervised models such as Logistic Regression, Random Forest, and Gradient Boosting in binary classification.
- Gained knowledge on feature engineering, normalization as well as preprocessing to maximize model performance and generalization.
- Learned to analyze the models based on such metrics as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.
- Put in place a full ML pipeline and used it in a real-time fraud prediction model based on a Streamlit web application.
- Better skills in the interpretation of prediction outcomes and meaningful risk explanations to facilitate financial decision-making.

IX. CONCLUSION

The objective of this work is to propose an efficient and intelligent credit card fraud detection method with the machine learning techniques. It considers transaction behavior analysis, supported by effective pre-processing methods that are able to highlight suspicious activities coming from big data, which may otherwise go unnoticed. In fact, three different models have been used in this work: Logistic Regression, Random Forest, and Gradient Boosting. This allows for thorough comparisons of performances to be made hence assuring that the final system will be both accurate and reliable. The prediction capability was further enhanced with the ensemble-based strategy, combining all strengths of the models.

Integrating the trained model into a Streamlit web application allows for practical and user-friendly deployment in real-time fraud analysis, thus enabling the usage of real-time accurate fraud alerts by financial institutions and security analysts or even customers. This

will make the proposed system more scalable and flexible toward improving the financial cybersecurity ecosystem and reducing economic losses due to fraudulent transactions in general. Improvements may be achieved by including some deep learning techniques, continuous model training, and incorporation of more contextual features that could further improve accuracy and robustness against evolving fraud patterns.

X. REFERNCES

[1] Dal Pozzolo, A., Caelen, O., Waterschoot, S., & Bontempi, G. Learned lessons in credit card fraud detection from a practitioner perspective. IEEE World Congress on Computational Intelligence, 2015.

[2] Friedman, J. H. Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232, 2001.

[3] Whitrow, C., Hand, D. J., Juszczak, P., et al. (2009). "Transaction aggregation as a strategy for credit card fraud detection." Data Mining and Knowledge Discovery, 18(1), 30-55.

[4] Maes, S., Tuyls, K., Vanschoenwinkel, B. (2002). "Credit card fraud detection using Bayesian and neural networks." First International NAISO Congress on Neuro Fuzzy Technologies.

[5] Breiman, L. Random Forests. Machine Learning Journal, 45(1), 5–32, 2001.

[6] Streamlit Team. Streamlit Documentation. Streamlit Inc. (Accessed 2025).

[7] S. Wang, M. Li, and H. Zhang, “Machine learning in fraud detection:A systematic literature review,” ACM Computing Surveys, vol. 54, no.2, pp. 1–36, 2021