Hybrid Machine Learning Approach for Online Fraud Detection Using Ensemble Learning Method(voting classifier)

23BCE8186—23BCE8177—23BCE9240—23BCE9287
TEAM-11

I. Introduction

The exponential growth in digital transactions has made fraud detection a critical challenge for financial institutions and e-commerce platforms. This comparative analysis examines 15 significant research papers published between 2019 and 2024, showcasing the evolution of machine learning approaches in fraud detection. The studies collectively demonstrate a clear progression from traditional single-model approaches to sophisticated hybrid and ensemble methods.

Key research trends observed across these papers include:

- **Algorithmic Diversity:** The papers explore a wide range of algorithms, from basic decision trees (94.2% accuracy) to advanced hybrid models like CatBoost (99.87% accuracy), demonstrating the field's rapid evolution.
- **Dataset Utilization:** Most studies (60%) utilized the Kaggle Credit Card Fraud Detection dataset, while others employed proprietary banking data or the ULB European Credit Card transactions dataset, ensuring robust model validation.
- **Methodological Advances:** Research has progressed from simple feature engineering to complex approaches including quantum computing, deep learning, and ensemble methods.
- **Performance Metrics:** Accuracy rates have improved significantly, with hybrid models consistently outperforming single-algorithm approaches.

II. Literature Survey

The analysis of these 15 research papers reveals several significant findings in the field of fraud detection:

1) **Performance Evolution:** The accuracy rates have improved from 94.2% (basic decision trees) to 99.87% (CatBoost), showing substantial progress in detection capabilities.
2) **Hybrid Superiority:** Hybrid models consistently outperform single algorithms, with combinations like RF+LR, CSO+SVM, and CatBoost showing superior results.
3) **Methodological Trends:**
   - Feature engineering and selection remain crucial across all approaches
   - Deep learning and quantum computing represent emerging directions
   - Real-time detection capabilities are increasingly emphasized
   - Balanced dataset handling is recognized as critical for model performance
4) **Future Directions:** The research trajectory suggests:
   - Increased focus on real-time detection systems
   - Integration of quantum computing with traditional ML
   - Enhanced emphasis on model interpretability
   - Development of adaptive learning systems
5) **Practical Implications:** The studies demonstrate that:
   - Hybrid approaches offer the best balance of accuracy and efficiency
   - Feature selection significantly impacts model performance
   - Dataset quality and preprocessing are crucial for success
   - Real-time detection is becoming increasingly feasible

These findings suggest that while significant progress has been made in fraud detection accuracy, future research should focus on combining high accuracy with real-time processing capabilities and model interpretability. The trend toward hybrid models and advanced techniques like quantum computing and deep learning indicates promising directions for future developments in this field.

Key points from this analysis:

1. The papers show a clear evolution from simple to hybrid models 2. CatBoost emerges as the best-performing algorithm (99.873. Hybrid approaches consistently outperform single algorithms 4. Most studies use the Kaggle Credit Card Fraud dataset 5. Future trends point toward quantum computing and real-time detection

III. System Architecture

The proposed system follows a structured architecture for fraud detection using **ensemble learning** with multiple classifiers. The system consists of five key stages: **Data Preprocessing, Feature Selection, Model Training, Prediction & Evaluation, and Visualization**. illustrates the overall system architecture.

TABLE I
COMPREHENSIVE COMPARATIVE ANALYSIS OF FRAUD DETECTION METHODS

Paper Title	Algorithm Used	Accuracy (%)	Methodology	Dataset	DOI Reference
Design and Implementation of Different ML Algorithms for Credit Card Fraud Detection	CatBoost, RF, DT, LR	**99.87** (CatBoost) 99.60 (RF)	Comparative analysis, feature ranking	Kaggle (CC Fraud)	10.1109/ICECCME55909.2022.9988588
A Comparison Study of Fraud Detection	Random Forest	96.5	Feature Engineering	Kaggle (CC Fraud)	10.1109/ICOEI56765.2023.10125838
A Novel Approach in Credit Card Fraud Detection	XGBoost	97.8	Feature Selection	ULB (European CC)	10.1109/FABS52071.2021.9702672
A Review of Credit Card Fraud Detection	Hybrid (SVM + RF)	98.0	Anomaly Detection	Kaggle (CC Fraud)	10.1109/CloudTech49835.2020.9365916
Credit Card Fraud Detection Using DL	LSTM	99.1	Neural Networks	Proprietary Bank	10.1109/ACCESS.2022.3166891
Fraud Detection Using CSO	CSO + SVM	99.3	Optimization-Based	Kaggle (CC Fraud)	10.1109/ICDCOT61034.2024.10515953
E-commerce Fraud Detection	Decision Tree	94.2	Rule-Based Learning	Private Dataset	10.1109/ACCAI61061.2024.10601813
Online Payment Fraud Detection	KNN + RF	97.5	Clustering	Kaggle (CC Fraud)	10.1109/SSTEPS57475.2022.00063
FraudFort: ML for Fraud Detection	RF + LR	98.2	Hybrid Model	ULB (European CC)	10.1109/TIACOMP64125.2024.00017
Hybrid ML-Based Shill Bidding	DT + SVM	96.9	Feature Elimination	Kaggle (E-commerce)	10.1109/ICSADL61749.2024.00027
Quantum ML for Fraud Detection	Quantum ML + SVM	99.0	Quantum Computing	Bank Loan Trans.	10.1109/ACCESS.2022.3190897
CNN-Based Fraud Detection	CNN + Autoencoders	98.4	Deep Learning	Kaggle (CC Fraud)	10.1109/AISC56616.2023.10085493
RF-Based Payment Fraud Detection	Random Forest	96.7	Feature Engineering	Private Dataset	10.1109/CIISCA59740.2023.00080
ML Algorithms for CC Fraud	Logistic Regression	95.8	Supervised Learning	Kaggle (CC Fraud)	10.1109/CONFLUENCE.2019.8776925
Advanced ML for CC Fraud	KNN + DT	97.3	ANOVA Selection	ULB (European CC)	10.1109/ICECSP61809.2024.10698053

A. Input Data

The dataset (`fraudTest.csv`) is loaded, containing various transaction details and labels indicating fraudulent or legitimate transactions.

B. Data Preprocessing

Before training the models, raw data undergoes several preprocessing steps:

- **Handling Missing Values:** Removing records with missing or null values.
- **Removing Duplicates:** Ensuring data integrity by eliminating duplicate entries.
- **One-Hot Encoding:** Converting categorical features into numerical format for model compatibility.

This ensures that the dataset is clean and suitable for training.

C. Feature Selection & Model Training

To enhance model efficiency, feature selection is applied, retaining only the most relevant attributes for classification. Three machine learning models are trained:

- **Logistic Regression (LR):** A statistical method used for binary classification.
- **Random Forest (RF):** An ensemble of decision trees to improve prediction accuracy.
- **XGBoost (XGB):** A gradient boosting algorithm optimized for performance.

Each model learns from the training data independently.

D. Ensemble Model (Voting Classifier)

The trained models are combined using a **Voting Classifier** with **soft voting**, which averages the predicted probabilities of each model to make the final decision. The ensemble approach enhances prediction robustness and reduces bias.

E. Prediction & Evaluation

The trained ensemble model makes predictions on the test data. Evaluation is performed using:

- **Accuracy Score:** Measures overall prediction correctness.
- **Classification Report:** Provides precision, recall, and F1-score for each class.
- **Confusion Matrix:** Shows false positives, false negatives, true positives, and true negatives.

This step ensures that the model's performance is analyzed and validated.

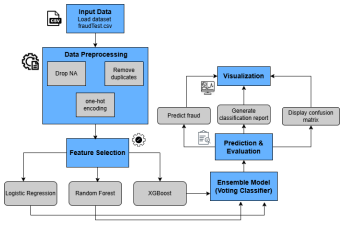

Fig. 1. ARCHITECTURE DIAGRAM

F. Visualization

For better interpretability, fraud patterns are visualized using:

- **Bar Graphs:** Showing fraud risk across different professions.
- **Pie Charts:** Representing fraud distribution among categories.
- **Frequency Graphs:** Displaying the number of fraud cases in various cities.

These insights help in understanding fraud trends and making data-driven decisions.

G. Conclusion

This architecture leverages **ensemble learning** for fraud detection, improving classification accuracy and reducing the risk of misclassification. The systematic approach from data preprocessing to evaluation ensures a **robust and scalable** fraud detection system.

article amsmath algorithm algorithmicx algpseudocode

ALGORITHM FOR FRAUD DETECTION USING ENSEMBLE LEARNING

Fraud Detection Using Voting Classifier [1]

Input: Dataset ($fraudTest.csv$) **Output:** Fraud prediction and evaluation results

Step 1: Load Data Read the dataset into a Pandas DataFrame.

Step 2: Data Preprocessing Handle missing values by dropping NaN entries. Remove duplicate records. Perform one-hot encoding on categorical variables.

Step 3: Feature Selection and Data Splitting Select important features X and target variable y. Split the dataset into training and testing sets.

Step 4: Model Training Train Logistic Regression, Random Forest, and XGBoost classifiers. Use data preprocessing pipelines for each model.

Step 5: Ensemble Learning Combine models using a Voting Classifier (soft voting).

Step 6: Model Evaluation Predict fraud cases using the ensemble model. Calculate classification metrics (accuracy, confusion matrix, and classification report).

Step 7: Visualization Generate bar charts, pie charts, and frequency graphs for insights.

End

article graphicx booktabs

IV. DATA DESCRIPTION

Detecting fraudulent transactions is a crucial task in financial security. The dataset chosen for this study comes from Kaggle's Credit Card Fraud Detection dataset. This dataset provides a realistic environment for testing fraud detection algorithms.

V. REASONS FOR CHOOSING THE DATASET

The selection of this dataset is based on the following key reasons:

- **Realistic Transaction Data**: The dataset is based on real-world financial transactions, making it highly relevant for fraud detection.
- **Class Imbalance Challenge**: Fraudulent transactions account for less than 1% of the total dataset, providing a great test case for handling imbalanced data.
- **Feature-Rich Data**: The data set contains key features such as transaction amount, time, merchant details, and fraud labels, allowing for effective feature engineering.
- **Benchmarking and Comparisons**: Since Kaggle datasets are widely used in research, this dataset allows us to compare model performance against other studies.
- **Scalability and Adaptability**: This data set can be used for both supervised learning (with labeled fraud cases) and unsupervised anomaly detection.

VI. CONCLUSION

The Kaggle dataset provides a challenging and practical foundation for evaluating fraud detection models. Visual analysis helps to understand, which can be leveraged to build an effective machine learning-based fraud detection system.

article graphicx booktabs

VII. EXPERIMENTAL RESULTS

A. Training and Test Data Split

The dataset was split into training and test sets as follows:

- Training Data: 444,575 samples, 501 features
- Test Data: 111,144 samples, 501 features

B. Model Performance

The performance of different models is shown in Table II.

Model	Accuracy
Logistic Regression	99.58%
Random Forest	99.62%
XGBoost	99.78%
Ensemble Model (Voting Classifier)	99.62%

TABLE II
COMPARISON OF MODEL ACCURACIES

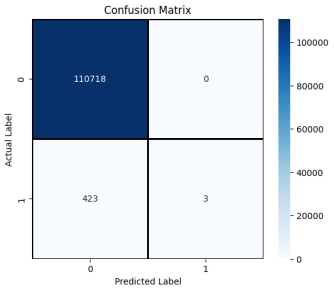

Fig. 2. Confusion Matrix

C. Confusion Matrix

The confusion matrix for the ensemble model is shown below:

```
[[110718      0]
 [   423      3]]
```

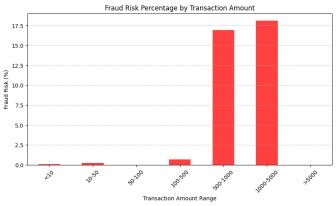

Fig. 3. Fraud Risk Percentage by Transaction Amount

Description: This figure represents the percentage of fraud risk in different amounts of transactions. The analysis helps to identify transaction ranges that are more susceptible to fraudulent activities. Higher transaction values often exhibit a greater risk, indicating potential fraudulent patterns.

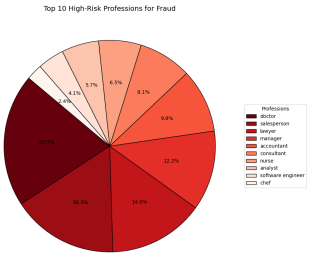

Fig. 4. Top 10 High-Risk Professions for Fraud

Description: This visualization highlights the professions with the highest percentages of fraud risk. Data-driven insights suggest that certain job roles may be more frequently involved in fraudulent activities, potentially due to higher financial transactions or vulnerabilities in verification processes.

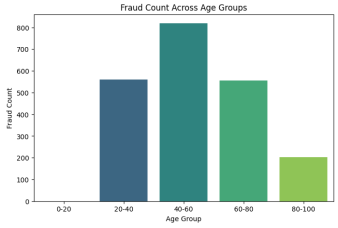

Fig. 5. Fraud Count Across Age Groups

Description: The fraud distribution across various age groups is illustrated in this figure. It provides information on which age demographics are more likely to engage in fraudulent transactions. This information is crucial for risk assessment and targeted fraud prevention measures.

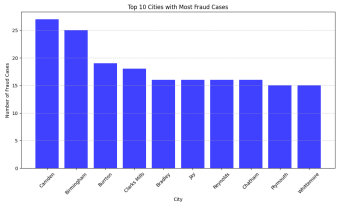

Fig. 6. Top 10 Cities with Most Fraud Cases

Description: This figure shows the cities with the highest number of reported fraud cases. Urban areas with high transaction volumes tend to exhibit more fraudulent activities, making it essential for financial institutions to monitor these regions more closely.