Final Project Milestone 1

Team: Nadezhda Shiroglazova, Prathamesh Nehete, Sheng Hu

Abstract

Financial institutions and consumers are seriously threatened by credit card fraud, which causes billions of dollars' worth of losses every year worldwide. This study tackles the problem of employing machine learning techniques to identify fraudulent credit card transactions on a highly unbalanced dataset in which fraud accounts for only 0.172% of all transactions. We apply and compare supervised learning techniques to maximize fraud detection performance while reducing false positives using the Kaggle Credit Card Fraud Detection dataset, which comprises 284,807 transactions from European cardholders in September 2013.

Our approach focuses on improving model learning from the minority fraud class by addressing class imbalance through three resampling techniques: undersampling, SMOTE oversampling, and hybrid sampling. We assess three main supervised models (Random Forest, XGBoost, and Logistic Regression) that represent various algorithm families and have demonstrated good performance in earlier fraud detection studies. To complement the PCA-transformed features in the dataset, we engineer fundamental behavioral features such as temporal patterns and amount deviations.

In order to reflect operational priorities, our evaluation uses realistic testing protocols with the original imbalanced test set and emphasizes recall, precision, and F1-score over accuracy. We provide error analysis to comprehend model failure modes and perform threshold optimization to balance false positives against false negatives. This strategy seeks to demonstrate practical applicability for fraud detection systems by achieving detection rates above 80% while maintaining precision above 85%. Our results set performance standards for this popular dataset and offer insights into practical methods for addressing highly imbalanced fraud detection issues.

Problem Statement

The primary issue is how uncommon fraudulent transactions are in real-world statistics. Just 492 of the 284,807 transactions in the sample are fraudulent, resulting in a class imbalance ratio of roughly 1:578. Because of this extreme imbalance, traditional machine learning algorithms are able to forecast all transactions as authentic while utterly failing to detect actual fraud, resulting in falsely high accuracy above 99%. In this situation, traditional accuracy measures become worthless because a model that never reports a transaction as fraudulent would still appear to be 99.8% accurate despite offering no value for preventing fraud.

Financial organisations must deal with excessive false positive rates in which valid transactions are mistakenly reported as fraudulent. This leads to customer annoyance and rejected transactions and irate clients and possibly lost revenue. Up to 95% of alerts in poorly calibrated systems are false positives according to research which can overburden fraud investigation teams. On the other hand low detection rates lead to immediate financial losses and legal liabilities and regulatory penalties. Recent projections indicate that global losses from credit card fraud exceeded 32 billion dollars in 2023 and could increase to 43 billion dollars by 2026 if effective measures are not implemented (Marazqah Btoush et al., 2023).

The challenge of class imbalance has been extensively documented in the literature. A systematic review of 181 journal articles found that this remains the most significant obstacle in fraud detection research with the Kaggle credit card dataset serving as the most commonly used benchmark for evaluating solutions (Marazqah Btoush et al., 2023). The review revealed that 74 percent of studies relied on supervised learning techniques which require labeled fraud data that is inherently scarce and expensive to obtain. This heavy dependence on supervised approaches presents additional challenges because fraudsters constantly adapt their strategies making historical fraud patterns less reliable for detecting new attack methods.

Traditional machine learning models face further difficulties when evaluated under realistic conditions. Research by Popova and Gardi (2024) demonstrated that models trained on balanced datasets through undersampling produce deceptively optimistic results during testing. When these same models were evaluated on the original imbalanced test set the precision dropped dramatically. For example Random Forest precision fell from high values to just 8.31 percent and Logistic Regression dropped to 5.01 percent. This finding highlights the critical importance of practical evaluation procedures that reflect real world deployment conditions rather than artificial laboratory settings.

The problem extends beyond simple class imbalance to include concept drift where fraud patterns evolve over time as criminals develop new techniques to evade detection systems. Mienye and Jere (2024) identified concept drift as a major challenge noting that static models trained on historical data become increasingly ineffective as fraudulent behavior changes. Additionally the high dimensionality of transaction data and the need for real time processing capabilities add computational complexity to an already difficult problem. Financial institutions require systems that can process thousands of transactions per second while maintaining both high detection rates and low false positive rates.

Current approaches also struggle with the tradeoff between detection sensitivity and operational practicality. Fariha et al. (2025) noted that even a 5 percent false positive rate corresponds to approximately 14,240 false alarms on the standard dataset which creates an enormous burden for manual review teams. The operational cost of investigating false positives can exceed the losses from undetected fraud in some cases. This creates a complex optimization problem where improving fraud detection must be balanced against the resources required to handle the resulting alerts.

Research Objectives

Based on the identified problems and gaps in current literature this project aims to achieve the following objectives:

Develop and compare at least four machine learning and deep learning models for credit card fraud detection including traditional classifiers like Random Forest and XGBoost and advanced techniques like neural networks and autoencoders

Implement and evaluate multiple resampling strategies including undersampling and oversampling and hybrid approaches to determine the most effective method for handling severe class imbalance

Establish appropriate evaluation metrics beyond accuracy including precision and recall and F1 score and AUC ROC that provide meaningful performance assessment for imbalanced classification problems

Create an automated end to end machine learning pipeline that facilitates data collection and feature engineering and model development and deployment following the CRISP-DM methodology

Provide actionable recommendations for financial institutions based on the tradeoff analysis between detection rates and false positive rates with quantified business impact in terms of potential fraud losses prevented

Alignment with Project Timeline

Week 3 (Current) Define scope and finalize problem statement and collect annotated bibliography of 6 peer reviewed articles establishing the theoretical foundation for our methodology choices

Week 4: Finalize data sources using the Kaggle credit card fraud dataset and begin preliminary data exploration

Week 6: Complete feature engineering and exploratory data analysis including behavioral features like transaction amount deviation and temporal patterns as suggested by Fariha et al. (2025)

Week 8: Design and develop minimum four models and evaluate using appropriate metrics. Models will include Random Forest and XGBoost based on their strong performance documented in Popova and Gardi (2024) and deep learning approaches like autoencoders based on recommendations from Mienye and Jere (2024)

Week 10: Final presentation with findings and recommendations and lessons learned and deployment strategy

Annotated Bibliography

Paper 1: Credit Card Fraud Detection

Popova, Iva, and Hamza A. A. Gardi. “Credit Card Fraud Detection.” Paper presented at **ETIT-KIT**, Germany, 2024.

Summary (Nadezhda Shiroglazova):

Study in this paper evaluates 5 machine learning models - Logistic Regression, Random Forest, XGBoost, K-Nearest Neighbors, and Multi-Layer Perceptron for credit card fraud detection using three resampling strategies to deal with severe class imbalance. Since fraud rate is only 0.172%. The work shows undersampling, oversampling, and a hybrid approach that combines both techniques, and all models are evaluated on initial dataset to imitate real data.

They uses same credit card dataset we are interested in, containing 284,807 transactions over two days in September 2013, with 28 PCA-transformed features plus Time and Amount variables. The paper's key feature is the hybrid sampling approach, which increases fraudulent transactions to approximately 2-2.5% in the training data while maintaining a higher proportion of legitimate transactions. This really improves model learning without data loss.

The results show that hybrid sampling greatly improves recall for all models while keeping precision at an acceptable level. The Random Forest model did the best overall on the imbalanced test set, with 94.2% precision, 78.9% recall, and 85.8% F1-score. The hybrid method helped models that were sensitive to class distribution the most. For example, Logistic Regression's recall went from 63.4% to 78.1% (+23.2%) and MLP's recall went from 65.9% to 81.3% (+23.4%). Tree-based models (Random Forest and XGBoost) showed more modest improvements, as they performed well even on imbalanced data.

The study highlights that testing on balanced datasets yields results that are deceptively optimistic. Precision significantly decreased when models trained on undersampled data were assessed on the initial unbalanced test set (e.g., Random Forest: 8.31%, Logistic Regression: 5.01%), underscoring the crucial significance of practical evaluation procedures. The authors come to the conclusion that hybrid sampling provides the best compromise between practical deployability and detection sensitivity, improving fraud detection without significantly compromising accuracy or overloading analysts with false positives.

Paper 2:Advanced Fraud Detection Using Machine Learning Models: Enhancing Financial Transaction Security

Fariha, Nudrat, Md Nazmuddin Moin Khan, Md Iqbal Hossain, Syed Ali Reza, Joy Chakra Bortty, Kazi Sharmin Sultana, Md Shadidur Islam Jawad, Saniah Safat, Md Abdul Ahad, and Maksuda Begum. "Advanced Fraud Detection Using Machine Learning Models: Enhancing Financial Transaction Security." *International Journal of Accounting and Economics Studies* 12, no. 2 (2025): 85-104.

Summary (Nadezhda Shiroglazova):

In order to identify fraudulent patterns without the need for labeled training data, this paper presents a comprehensive unsupervised machine learning framework for credit card fraud detection that tackles the crucial issue of label scarcity by utilizing multiple anomaly detection techniques (Isolation Forest, One-Class SVM, and deep autoencoders) in conjunction with clustering techniques (K-Means and DBSCAN). The Adaptive Risk-Scoring Framework (ARF), which uses online learning with mini-batch gradient descent to dynamically calibrate detector weights based on contextual factors like demographic patterns, merchant category volatility, and regulatory environments, is the main innovation of the authors.

Beyond the basic PCA-transformed features in the Kaggle dataset, the framework incorporates behavioral feature engineering to create metrics like amount deviation from personal averages, inter-transaction time intervals, merchant category frequency patterns, and temporal indicators (hour of day, day of week, weekend flags). The study reports that all models achieved about 95% detection rates with 5% false positive rates, with the autoencoder achieving the highest AUC of 0.971. Analysis shows that food and beverage merchants (especially bars and food trucks) exhibit the greatest vulnerability, while nighttime transactions show the highest risk despite lower volume. The study details practical implementations with 30–70% fewer false positives and 35–50% fewer fraud losses across major payment platforms, such as Stripe, Visa, and Mastercard.

Nevertheless, the paper has serious methodological flaws that seriously compromise its purported contributions. Most importantly, it is impossible to determine whether the 1-2% of transactions that are flagged are actually fraudulent or just unusual but legal activity because all performance metrics are assumption-based and lack verified fraud labels. The reported 5% false positive rate corresponds to about 14,240 false alarms on the dataset, but the paper does not sufficiently address the operational burden of manual review and the ensuing customer friction. However, this paper shows the value of multi-method consensus in anomaly detection and offers a good template for feature engineering for our project. However, we should simplify the ARF complexity, validate all approaches against labeled fraud data, and perform appropriate threshold sensitivity analysis.

Paper 3: Systematic Review of ML/DL Techniques

Marazqah Btoush, Eyad Abdel Latif, Xujuan Zhou, Raj Gururajan, Ka Ching Chan, Rohan Genrich, and Prema Sankaran. "A Systematic Review of Literature on Credit Card Cyber Fraud Detection Using Machine and Deep Learning." PeerJ Computer Science 9 (2023): e1278.

Summary (Prathamesh Nehete):

This systematic review examines 181 peer reviewed journal articles published between 2019 and 2021 that focus on credit card cyber fraud detection using machine learning and deep learning techniques. The authors conducted searches across six major digital libraries including Google Scholar and IEEE Xplore and Scopus and Web of Science and ACM and SpringerLink. They used a rigorous methodology with clearly defined inclusion and exclusion criteria to identify the most relevant studies in this field.

The review addresses four key research questions about the ML/DL techniques used for fraud detection and the proportion of supervised versus unsupervised approaches and overall model performance and future research directions. The findings reveal that Random Forest is the most frequently used ML technique appearing in 74 articles followed by Logistic Regression in 52 articles and Support Vector Machine in 56 articles. Among deep learning approaches the Artificial Neural Network appeared in 36 articles while Autoencoders were used in 18 articles and LSTM networks appeared in 8 articles.

A significant finding is that 74 percent of reviewed articles used supervised learning techniques while only 12 percent used unsupervised methods and another 12 percent combined both approaches. Semi supervised learning was used in just 2 percent of studies. The authors note that this heavy reliance on supervised learning presents challenges because labeled fraud data is scarce and fraudsters constantly change their behavior patterns. The review also identifies the Kaggle credit card dataset as the most commonly used benchmark containing 284807 transactions with only 492 fraudulent cases representing a severe class imbalance of 0.172 percent.

For our project this paper provides an excellent foundation for understanding the landscape of fraud detection techniques. It confirms that our planned approach using multiple models including Random Forest and XGBoost and neural networks aligns with current best practices. The review also highlights the importance of addressing class imbalance through techniques like SMOTE which we plan to implement. However the authors recommend that future research should explore unsupervised and semi supervised methods more extensively which suggests an opportunity for our project to incorporate anomaly detection approaches like autoencoders to strengthen our methodology.

Paper 4: Deep Learning Review for Credit Card Fraud Detection

Mienye, Ibomoiye Domor, and Nobert Jere. "Deep Learning for Credit Card Fraud Detection: A Review of Algorithms, Challenges, and Solutions." IEEE Access 12 (2024): 96893-96910.

Summary (Prathamesh Nehete):

This IEEE Access review paper provides a comprehensive examination of deep learning techniques specifically applied to credit card fraud detection. The authors present detailed descriptions of widely used DL architectures including Convolutional Neural Networks and Recurrent Neural Networks and Long Short Term Memory networks and Gated Recurrent Units. The paper goes beyond simple technique descriptions by analyzing the unique challenges that arise when applying these methods to fraud detection problems.

The review identifies several critical challenges in fraud detection systems. Class imbalance remains the primary obstacle since fraudulent transactions typically represent less than 1 percent of all transactions. The authors discuss various solutions including resampling techniques like SMOTE and ADASYN as well as cost sensitive learning approaches that assign higher misclassification costs to the minority fraud class. Concept drift is another major challenge where fraud patterns evolve over time as criminals adapt their strategies to avoid detection. The paper also addresses the high dimensionality of transaction data and the need for real time processing capabilities.

A key contribution of this paper is its performance comparison across different DL architectures. The authors report that LSTM networks excel at capturing temporal dependencies in transaction sequences which is valuable for detecting patterns over time. Autoencoders trained only on legitimate transactions can effectively identify anomalies by measuring reconstruction error. CNN architectures have shown promise in extracting spatial features from transaction data when properly formatted. The paper notes that ensemble approaches combining multiple DL models often achieve the best results with some studies reporting accuracy above 99 percent and AUC scores exceeding 0.98.

This paper is highly relevant to our project because it provides specific guidance on implementing deep learning models for fraud detection. The discussion of autoencoders as anomaly detectors directly supports our plan to include this technique in our model comparison. The paper also emphasizes the importance of using appropriate evaluation metrics beyond accuracy such as precision and recall and F1 score and AUC ROC which is critical given the imbalanced nature of fraud datasets. The authors recommend that practitioners should focus on reducing false positives while maintaining high fraud detection rates which aligns with our goal of creating a practically deployable system. Their suggestion to combine multiple approaches in an ensemble framework provides a clear direction for our model development strategy.

Paper 5: Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms

Alarfaj, F. K., I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed. “Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms.” *IEEE Access* 10 (2022): 39700–39715. https://doi.org/10.1109/ACCESS.2022.3166891.

Summary (Sheng Hu):

In this article, Authors focus on the detection of credit card fraud in online transactions, considering it a high-impact issue as fraud causes significant losses to both cardholders and financial institutions. The authors highlight several practical obstacles that make this task difficult in real-world operations: limited access to public fraud data, severe class imbalance where fraudulent transactions are extremely rare compared to legitimate ones, evolving fraud tactics over time, and the high false positive rate of bank risk control systems, which can frustrate customers and waste investigators' time. To make their research foundational, they compare it with widely used machine learning methods in the literature such as decision trees, random forests, support vector machines, logistic regression, and XGBoost, arguing that despite progress, low accuracy and operational costs prompt the need to apply more modern deep learning techniques to reduce fraud losses.

From a methodological perspective, the paper presents a comparative empirical study that evaluates machine learning and deep learning methods on the European Credit Card benchmark dataset. Their experimental process involves first applying machine learning algorithms, which they claim improve fraud detection performance to some extent, and then applying three convolutional neural network (CNN) architectures aimed at extracting useful patterns from transaction data and improving classification results. A key design lever they explore is model capacity: they report that adding more layers to the CNN architectures can further enhance detection accuracy, and they also achieve broader model variations by changing training settings such as the number of hidden layers and training cycles. Regarding the reported results, the paper emphasizes that the best-case metrics are "approximately 99.9% accuracy, 85.71% F1 score, 93% precision, and 98% AUC, and concludes that the proposed CNN-based methods outperform the evaluated state-of-the-art machine learning and deep learning benchmarks on this dataset." Importantly, they also conduct additional experiments to balance the data and apply deep learning to reduce the false negative rate of fraud detection, as false negatives can cause particularly severe losses in actual deployment. Overall, the contribution lies in the parallel comparison of machine learning and deep learning, the experimentation with CNN architectures, and the evaluation focused on imbalanced data. It clearly indicates that these methods can be effectively applied to real-world fraud detection.

Paper 6: Credit Card Fraud Data Analysis and Prediction Using Machine Learning Algorithms

Karunya R. V., Samyuktha Ganesh, Muralidharan D., Brindha G. R., and Muthu Thiruvengadam. "Credit Card Fraud Data Analysis and Prediction Using Machine Learning Algorithms." *Security and Privacy* 8 (2025): e70043. https://doi.org/10.1002/spy2.70043.

Summary (Sheng Hu):

In this article, the author delves into the issue of credit card fraud detection in the context of the rapid growth of digital payments, where banks need models capable of flagging suspicious transactions at point-of-sale terminals and online platforms. The paper highlights a practical workflow that begins with exploratory data analysis (EDA) to understand customer/transaction behavior and data quality, including detecting duplicates and outliers, applying feature encoding and scaling, performing dataset balancing to mitigate class imbalance, and using charts and visualizations to extract insights, before moving on to build predictive models.

In terms of modeling, the author compares several machine learning methods, explicitly mentioning logistic regression, k-nearest neighbors (kNN), and support vector machines (SVM), and then proposes a probability-based variant of kNN. Unlike conventional approaches, this method does not rely on the typical kNN distance metric in the feature space but instead uses the probability values generated by logistic regression to drive neighborhood-based classification decisions, and aiming to improve kNN's performance on imbalanced fraud datasets while maintaining relatively light computational requirements. The study positions this hybrid approach as a way to leverage the calibration or scoring output of logistic regression while retaining the intuitive local decision-making nature of kNN, and evaluates it in the broader context of imbalance-aware preprocessing and baseline comparisons.