UDC 004
IRSTI 28.23.37

https://doi.org/10.55452/1998-6688-2026-23-2-242-249

[1]*Azizov T.,
Master's student, ORCID ID: 0009-0008-3923-6746,
*e-mail: timazizov@gmail.com
[2]Abdiakhmetova Z.,
Associate Professor, ORCID ID: 0000-0001-8702-511X,
e-mail: zukhra.abdiakhmetova@gmail.com
[1]Kazakh-British Technical University, Almaty, Kazakhstan
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan

AN SMOTE-BASED METHOD FOR ENHANCING CREDIT CARD FRAUD DETECTION

Abstract

Credit card fraud occurs most often in online purchases; therefore, it is crucial to employ better ways to find it to avoid financial loss. This paper discusses fraud detection by employing methods to generate synthetic data to improve detection models. We use the Kaggle credit card transaction dataset, implementing synthetic data generation using SMOTE as a way to balance the dataset, in which fraud cases comprise only 0.2% of cases, and perform feature engineering to better understand buying behavior. We experimented with five ML models–XGBoost, LightGBM, Random Forest, Neural Networks, and Logistic Regression; focusing on precision, recall, F1-score, and accuracy. The comparison indicates that XGBoost achieves its highest F1-score (82.57%) with good precision (93.75%) and recall (73.77%), indicating XGBoost can balance false positives and false negatives. Although all models performed with high accuracy (over 99.9%), this research focuses on highlighting precision and recall in fraud detection. The findings suggest that combining synthetic data with gradient-boosting algorithms can help fraud detection systems improve the security of online purchases.

Keywords: Credit card fraud, SMOTE, Machine learning, XGBoost, Imbalanced data, Online transactions.

Received November 11, 2025; revised February 19, 25, 2026; accepted March 3, 2026.

Introduction

The rapid development of online commerce has made digital transactions extremely convenient. Opening a bank account, ordering home supplies and groceries can be done from the comfort of one's couch. However, while this evolution in online transactions provides great comfort to most people, it also comes with various financial risks. This shift in digital payments has created opportunities for a plethora of businesses to scale, but it also has created opportunities for fraudsters and cybercriminals in the digital space. Fraud in banking typically refers to any unauthorized use of credit or debit cards in order to obtain money. Card cloning, card theft, or manipulation of online transactions without knowledge of the card holder are the main ways credit card fraud can be executed. Fraud in online purchases has been increasing in recent years, costing approximately $42 billion annually [1]. To make e-commerce a safer space, banks, payment providers, and other financial institutions must innovate, continuously developing the best possible methods for detecting suspicious activities.

Machine learning has established itself as a tool that can help mitigate fraud in financial transactions. It performs better in detecting suspicious activities than a traditional rule-based system, though it requires more computational power [2]. Traditional rule-based systems performed worse because potential fraudulent actors can easily bypass fixed rules, whereas machine learning can evolve by learning different fraud detection patterns, which is not achievable with rigid rule-based systems. Many different approaches are used to find fraud in online payments. Rani and Mittal [3]

conducted an in-depth study of AI systems that examined money transfers in real time. The research showed that these systems, by leveraging machine learning algorithms, can spot unusual patterns somewhat quickly by learning and understanding what normal spending looks like for each person. If a certain transaction does not fit a particular pattern, the system flags it for review. That approach could help banks stop fraud before the client's money is gone. These systems show improvement over time as they are exposed to an increasing number of both fraudulent and non-fraudulent examples.

A particular research study by Peneti et al. [4] implemented the idea of pattern recognition and user behavior using ML models such as Logistic Regression and Random Forest achieving accuracy above 90%. This study highlighted the importance of data preprocessing, feature selection, and model evaluation. Statistical models in conjunction with machine learning algorithms can also be useful [5]. Research by Thar and Wai showed that combining Hidden Markov Models (HMM) and Gradient Boosting Classifiers (GBC) can achieve 96% accuracy with a recall of 89.38%.

Gradient boosting models were found to be highly successful in credit card fraud prediction due to their ability to handle complex, non-linear patterns of transaction information and excel with unbalanced datasets common in fraud discovery issues. In a study by Ahirwar [6], XGBoost achieved 99% accuracy and 82% F1 score. A further study by Ruchita et al. [7] compared XGBoost and LightGBM, and concluded that LightGBM actually outperformed XGBoost in accuracy, precision, recall and consequently, in F1 score.

Artificial Neural Networks (ANN) have also shown interesting results in detecting anomalies in transactions. Research by Anusha et al. [8] showed that ANN achieved 97.59% accuracy and 83.91% precision in detecting suspicious activities. Another study by Singh et al. [9] compared ANN with XGBoost and ANN achieved 4.2% higher accuracy. Unfortunately, this research did not explicitly provide information about the recall and precision of the two models, which is crucial when dealing with transaction data, since they may have imbalanced datasets. Overall, ANNs have the potential to be effective tools for credit card fraud detection, because of their ability to detect nonlinear patterns.

The importance of computational efficiency cannot be neglected when dealing with the detection of credit card fraud. Inefficient fraud detection models coupled with standard hardware can cause malicious transactions to pass through before being flagged. Khalid's [10] research concluded that 99.96% accuracy is achievable without compromising much computational efficiency. The developed model was able to handle up to 38 entries per second on fairly generic hardware. Financial institutions process a much greater number of transactions each second; however, this was done at a small scale – more powerful hardware coupled with horizontal scaling could achieve a much better result.

An important issue that has to be addressed is latency, which is one of the most crucial performance constraints for fraud detection systems and goes hand in hand with computational efficiency. In a modern digital payment world, it is vital for a model to detect the legitimacy of a transaction in a matter of milliseconds, so as to not disrupt customer flow and provide uninterrupted user experience. A study by Axenie et al. [11] demonstrated that while processing high-volume transaction streams, events can be processed in the 1 to 8 millisecond range. They used transaction data that simulated peak shopping periods such as Single's Day in China, processing around 2 million events at 40 kHz and a load of 40,000 events per second. Another experiment by Basani [12] revealed that the XGBoost model was able to handle 10,000 transactions per second, with approximately 7ms latency per transaction, which is consistent with performance metrics displayed in streaming PCA techniques applied to fraud detection.

Another major challenge, beyond efficiency issues, is dealing with imbalanced datasets coupled with the scarcity of publicly available data due to personal privacy concerns. To tackle these issues, synthetic data generation can be used [13, 14]. The authors implemented the PaySim tool in order to generate data that would mimic, often sensitive and private transactions. The result still yielded an imbalanced dataset, therefore to handle this issue, they used Conditional Generative Adversarial Networks (cGAN) and SMOTE+ENN (Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbors) for data augmentation and noise removal. The augmented data was then used to train different machine learning models such as Logistic Regression, Random Forest and

Support Vector Machines. The results show that using synthetic data significantly improved model performance, achieving high accuracy, precision, recall, and F1-score. This shows the effectiveness of GANs for data generation in cases with unbalanced datasets.

Materials and methods

The research used a Kaggle dataset that comprised real credit card transaction data from European cardholders in September 2013. The dataset exhibited a severe class imbalance, with over 280,000 transactions, of which only 492 were fraudulent. Most features (V1-V28), due to confidentiality constraints, were already pre-processed and transformed using Principal Component Analysis (PCA), having only 'Time' and 'Amount' variables remaining unchanged.

In order to improve model discriminative power, several feature engineering techniques were implemented. First, a new variable was introduced: 'Amount Time Ratio' to capture the relationship between transaction value and timing. Next, interaction terms between several selected features (e.g. V1-V4) were created to capture nonlinear relationships (Table 1). Lastly, to maintain temporal distribution patterns, time-based stratified splitting was implemented.

Table 1 – Kaggle Credit Card Dataset Illustrative Example

Time	V1	V2	V3	V4	Amount	Class
0	-1.50	0.80	-0.90	1.20	75.20	0
1	2.30	-1.10	0.60	-0.50	120.00	1
2	0.00	1.50	-1.30	0.30	45.75	0
3	-0.90	-0.40	1.20	-1.00	230.10	1
4	1.10	-1.60	0.90	0.00	90.90	0
5	-2.00	0.30	-0.70	0.60	310.50	1
6	0.70	1.20	-0.50	0.40	88.88	0
7	-1.30	-0.90	0.80	-0.70	150.00	1
8	1.50	0.60	-1.20	0.90	60.25	0
9	-0.20	-1.40	1.10	0.10	199.99	0

The dataset used for this research suffers from extreme class imbalance toward benign transactions, with less than 0.2% being problematic. Most datasets dealing with banking fraud exhibit similar issues; therefore, techniques to mitigate this problem need to be adapted. To address the issue of class imbalance, Synthetic Minority Over-Sampling Technique (SMOTE) was used. SMOTE generates synthetic data samples for the minority class– in our case, fraud – by averaging and mixing real ones, in order to make the class distribution more balanced. The SMOTE configuration was to over-sample the minority class until the target ratio of half the size of the majority class was reached. Conversely, majority class was under-sampled until it reached the target of a 1:2 ratio of minority to majority. The final step taken was to ensure that the generated data resembled real fraud cases. Data quality check was performed using Kernel Density Estimation (KDE) to compare the distribution of a certain feature in both (real and synthetic) cases, and synthetic data showed high statistical consistency with the original.

After the data is processed and target class balance is achieved, the training stage can begin. This research takes advantage of five different machine learning algorithms, due to their performance and the ability to find complex patterns in datasets. The models are Logistic Regression, Random Forest, XGBoost, LightGBM and Neural Network.

Logistic Regression: Logistic regression is a comparatively simple algorithm that is often used as a baseline. By using a weighted combination of input features, it predicts whether the transaction is fraudulent or not. Even though it expects a linear relationship between input and output, more often than not it can show high performance when data is well-processed.

Random Forest: Random Forest is an ensemble method that utilizes a combination of multiple decision trees. The final prediction of this algorithm is made by the average results of each decision tree.

XGBoost: XGBoost is a fairly popular algorithm in fraud detection, due to its high speed, accuracy and ability to handle missing and noisy data. It is a gradient boosting algorithm that builds decision trees sequentially.

LightGBM: LightGBM is another gradient boosting algorithm, but unlike XGBoost which splits trees level-wise, LightGBM splits trees leaf-wise. It is particularly powerful at handling large datasets that have a plethora of different features.

Neural Networks: Taking inspiration from the human brain, Neural Networks are powerful models capable of learning complex patterns in data. They consist of multiple layers of connected nodes that can detect subtle, hidden, and non-linear patterns, which can often be found in fraudulent transactions.

After synthetic data generation and model training, the next crucial step is to evaluate how accurate the models are. Several metrics have been employed to evaluate how effective models are in fraud detection:

Accuracy can be misleading when dealing with fraudulent data, for instance, with the dataset that we used for this research we can achieve as high as 99.8% accuracy without potentially detecting any fraud. However, it is still important to use it in conjunction to other metrics as a baseline. Accuracy can be defined using this formula:

$$ACCURACY = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1.1}$$

where TP (true positives) represents correctly identified fraudulent transactions, TN (true negatives) represents correctly identified benign transactions, FP (false positives) represents legitimate transactions wrongly detected as fraud, while FN (false negatives), conversely, represents fraudulent transactions that were missed by our model.

Precision in fraud detection measures the proportion of true fraud among all transactions that were flagged that way. It is crucial for a model to be able to detect

fraud with high precision, in order to minimize false flags and create customer inconvenience. Precision is defined as:

$$PRECISION = \frac{TP}{TP + FP}. \tag{1.2}$$

Recall (or sensitivity) could be considered a counterpart to precision. It measures the proportion of correctly identified frauds to all guesses. It is expressed as:

$$RECALL = \frac{TP}{TP + FN}. \tag{1.3}$$

F-1 score represents the harmonic mean between precision and recall. F-1 provides balance between these two metrics to evaluate model performance. This metric is particularly important for fraud detection as maximization of fraud detection (recall) and minimization of false flags (precision) both play a crucial role in the banking sector.

F1 can be calculated using this expression:

$$F1 = \frac{2 * PRECISION * RECALL}{PRECISION + RECALL}. \tag{1.4}$$

Together, these metrics outline the effectiveness of a particular model in properly detecting fraudulent transactions. It is vital to not only to measure the accuracy of the model, especially when dealing with imbalanced data, but also to look at metrics that assess precision and recall as well.

Results and discussion

Various algorithms for machine learning have been trained on the augmented set (original set along with synthetic fraud instances) and validated in a test set. The performance of each of them in total is given in Table 2.

Table 2 – Final model performance

Model	Accuracy	Precision	Recall	F1-Score
XGBoost	99.96%	93.75%	73.77%	82.57%
LightGBM	99.96%	90.00%	73.77%	81.08%
Neural Network	99.94%	100.00%	54.10%	70.21%
Logistic Regression	99.95%	93.02%	65.57%	76.92%
Random Forest	99.96%	97.62%	67.21%	79.61%

XGBoost performed the best overall with an F1-score of 82.57% with high precision (93.75%) also supplemented by strong recall (73.77%). This balance of precision over recall works most fittingly in fraud detection use cases where false positives as well as false negatives come at a very high cost. LightGBM was very close to XGBoost with F1-score of 81.08%, though with lower precision (90.00%) but with the same recall (73.77%). The similar behavior of these two gradient boosting libraries was in agreement with findings of the prior research [7].

Random Forest performed with the highest precision among all other models: it was accurate in classifying a transaction as fraudulent in 97.62% of all cases. This came at a cost to recall (67.21%), which resulted in it having an F1-score of 79.61%, placing it third in total. Neural Networks achieved perfect precision (100%) in terms of having no false positives, but at a cost of recall (54.10%), leading to a lower F1-score (70.21%). This type of trade-off would then be appropriate in cases where false alarms are very costly.

Logistic Regression, while being relatively simple compared to other models, performed admirably with an F1-score of 76.92%, which indicates the excellence of feature engineering task as well as synthetic data generation.

All models attained very high accuracy (more than 99.9%), which is to be expected since there was a natural class imbalance in fraud detection. This highlights the fact that precision, recall, and F1-score should be employed as important evaluation metrics for imbalanced classification issues.

Conclusion

Online payment transactions are increasing rapidly; hence robust and efficient methods are required to identify fraud in order to safeguard financial systems. The experiment demonstrates why it is necessary to use advanced techniques like synthetic data generation and adopt new machine learning strategies to correct class imbalance in credit fraud datasets. With the use of SMOTE to generate synthetic fraud samples and feature adjustments in order to gain insights into transaction behavior, we significantly enhanced model performance benchmarks. Out of all algorithms tested, XGBoost yielded the highest F1-score of 82.57%, high precision (93.75%), high recall (73.77%) while maintaining a good balance in lowering false positives and still enhancing fraud detection. Although all models possessed very high accuracy levels (over 99.9%), the experiment indicates that in unbalanced cases, precision as well as recall serve better to gauge effectiveness.

The experiment reveals a trade-off: Random Forest's highest precision (97.62%) reduces customer issues resulting from false alarms. Neural Networks produce flawless precision but lower recall. The experiment indicates that the choice between these two hinges upon what the company desires, either minimizing losses in terms of money (high recall) or maintaining customer trust (high precision).

Limitations include reliance upon artificial data that could fail to represent shifting fraud patterns, as well as conducting rapid testing in real-world scenarios. Future research could explore mixed-method approaches utilizing GANs to generate more accurate artificial datasets, learn in real time, as well as prioritize privacy. The research provides a benchmark for banks to make online purchasing secure in a cashless society through better fraud detection.

REFERENCES

1 Dorfleitner, G., and Jahnes, K. Banking fraud: Global financial impact and detection methodologies. Journal of Financial Crime, 29 (2), 456–471 (2022).

2 Khanum, A., Chaitra, K.S., Singh, B., and Gomathi, C. Fraud detection in financial transactions: A machine learning approach vs. rule-based systems. Proceedings of the 2nd International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2024 (2024).

3 Rani, S., and Mittal, A. Securing digital payments a comprehensive analysis of ai driven fraud detection with real time transaction monitoring and anomaly detection. Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2023, 2345–2349 (2023).

4 Peneti, S., Krishna, S.R., Kiran, A., and Tripathy, H.K. Credit card fraud detection using machine learning. Proceedings of 2nd International Conference on Advancements in Smart, Secure and Intelligent Computing, ASSIC 2024 (2024).

5 Thar, K.W., and Wai, T.T. Machine learning based predictive modelling for fraud detection in digital banking. Proceedings of the 21st IEEE International Conference on Computer Applications 2024, ICCA 2024, 249–253 (2024).

6 Ahirwar, N., Singh, D., and Maheshwar, K. Efficient credit card fraud detection based on multiple ml algorithms. 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), 1–7 (2024).

7 Ruchita, M., Bhargavi, M., Rakshita, M., Nandini, B.C., Aziz, I., and Gopi, J. Leveraging smote and random forest for improved credit card fraud detection. 2024 International Conference on Sustainable Communication Networks and Application (ICSCNA), 795–800 (2024).

8 Anusha, P., Bharath, S., Rajendran, N., Durga Devi, S., and Saravanakumar, S. Experimental evaluation of smart credit card fraud detection system using intelligent learning scheme. 2023 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES), 1–6 (2023).

9 Singh, A., Gill, K.S., Kumar, M., and Rawat, R. Beyond traditional methods: Evaluating advanced machine learning models for superior fraud detection. 2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS), 297–300 (2024).

10 Khalid, A.R., Owoh, N., Uthmani, O., Ashawa, M., Osamor, J., and Adejoh, J. Enhancing credit card fraud detection: An ensemble machine learning approach. Big Data and Cognitive Computing, 8 (1), 6 (2024).

11 Axenie, C., Tudoran, R., Bortoli, S., Al Hajj Hassan, M., Salort Sanchez, C., and Brasche, G. Dimensionality reduction for low-latency high-throughput fraud detection on datastreams. 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 1170–1177 (2019).

12 Reddy Basani, M.A. Data engineering and ml for real-time fraud detection in financial transactions. 2024 IEEE 8th Conference on Energy Internet and Energy System Integration (EI2), 91–96 (2024).

13 Mohapatra, A., Kumar, A., Kumar, B., Agarwal, H., and Priyadarshini, R. Synthetic data generation and handling data imbalance for mobile financial transactions. CSNT (2024).

14 Sultana, S., Rahman, M.S., and Afroj, M. An efficient fraud detection mechanism based on machine learning and blockchain technology. 2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT 2023, 162–168 (2023).

[1]*Азизов Т.,
магистрант, ORCID ID: 0009-0008-3923-6746,
*e-mail: timazizov@gmail.com
[2]Абдиахметова З.,
қауымдастырылған профессор, ORCID ID: 0000-0001-8702-511X,
e-mail: zukhra.abdiakhmetova@gmail.com

[1]Қазақстан-Британ техникалық университеті, Алматы қ., Қазақстан
[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы қ., Қазақстан

НЕСИЕЛІК КАРТАЛАРДАҒЫ АЛАЯҚТЫҚТЫ АНЫҚТАУДЫ ЖЕТІЛДІРУ ҮШІН SMOTE ӘДІСІНЕ НЕГІЗДЕЛГЕН ТӘСІЛ

Аңдатпа

Несиелік карталармен алаяқтық көбінесе онлайн-сатып алу кезінде орын алады, сондықтан қаржылық шығындардың алдын алу үшін оны анықтаудың тиімді әдістерін қолдану маңызды. Мақалада модельдердің сапасын арттыру мақсатында синтетикалық деректер генерациясын қолданатын алаяқтықты анықтау тәсілдері қарастырылады. Зерттеу барысында Kaggle несиелік карта транзакцияларының деректер жиынтығы пайдаланылып, SMOTE әдісі деректерді теңестіру үшін қолданылды, себебі алаяқтық жағдайларының үлесі барлығы 0,2% құрайды. Сатып алу мінез-құлқын жақсырақ түсіну үшін ерекшеліктерді инженерлік талдау жүргізілді. XGBoost, LightGBM, Random Forest, нейрондық желі және логистикалық регрессия сияқты бес машина оқыту моделі салыстырылды; дәлдік, еске түсіру, F1-мера және жалпы дәлдік көрсеткіштері бағаланды. Нәтижелер XGBoost моделі ең жоғары F1 көрсеткішін (82,57%) және жоғары дәлдік (93,75%) пен еске түсіру (73,77%) нәтижелерін көрсеткенін анықтады. Барлық модельдер жоғары дәлдікке (99,9%-дан жоғары) қол жеткізгенімен, зерттеу алаяқтықты анықтау кезіндегі дәлдік пен еске түсірудің маңыздылығын атап өтеді. Қорытындылай келе, синтетикалық деректер мен градиенттік бустинг алгоритмдерін біріктіру онлайн-төлемдер қауіпсіздігін арттыруға мүмкіндік береді.

Түйін сөздер: несиелік карта алаяқтығы, SMOTE, машиналық оқыту, XGBoost, теңгерімсіз деректер, онлайн-транзакциялар.

[1]*Азизов Т.,
магистрант, ORCID ID: 0009-0008-3923-6746,
*e-mail: timazizov@gmail.com
[2]Абдиахметова З.,
ассоциированный профессор, ORCID ID: 0000-0001-8702-511X,
e-mail: zukhra.abdiakhmetova@gmail.com

[1]Казахстанско-Британский технический университет, г. Алматы, Казахстан
[2]Казахский Национальный университет им. аль-Фараби, г. Алматы, Казахстан

МЕТОД НА ОСНОВЕ SMOTE ДЛЯ ПОВЫШЕНИЯ ЭФФЕКТИВНОСТИ ОБНАРУЖЕНИЯ МОШЕННИЧЕСТВА С КРЕДИТНЫМИ КАРТАМИ

Аннотация

Мошенничество с кредитными картами чаще встречается при онлайн-покупках, поэтому крайне важно использовать более эффективные способы его обнаружения, чтобы избежать финансовых потерь. В данной статье рассматривается выявление мошенничества с помощью методов генерации синтетических данных для улучшения моделей. Мы используем набор данных о транзакциях по кредитным картам Kaggle, реализуя генерацию синтетических данных с помощью SMOTE для балансировки набора данных, в котором случаи мошенничества составляют всего 0,2% случаев, и проводим разработку признаков для лучшего по-

нимания поведения покупателей. Мы экспериментируем с пятью моделями машинного обучения: XGBoost, LightGBM, Random Forest, Neural Networks и Logistic Regression, уделяя особое внимание точности, полноте, F1-оценке и достоверности. Сравнение показывает, что XGBoost достигает наивысшей F1-оценки (82,57%) при хорошей точности (93,75%) и полноте (73,77%), что свидетельствует о способности XGBoost сбалансировать ложноположительные и ложноотрицательные результаты. Хотя все модели показали высокую точность (более 99,9%), в данном исследовании основное внимание уделяется точности и полноте при выявлении мошенничества. Результаты показывают, что сочетание синтетических данных с алгоритмами градиентного усиления может помочь системам обнаружения мошенничества повысить безопасность онлайн-покупок.

Ключевые слова: мошенничество с кредитными картами, SMOTE, машинное обучение, XGBoost, несбалансированные данные, онлайн-транзакции.