International Journal of Integrative Research (IJIR)
Vol.3, No.7, 2025: 511-520

Comparison of Random Forest and K-Nearest Neighbors (KNN) Algorithms in Detecting Credit Card Transaction Anomalies

Fitria[1*], Emy Iryanie[2], Heldalina[3], Muhammad Syahid Pebriadi[4], Heru Kartika Candra[5]
Politeknik Negeri Banjarmasin
Corresponding Author: Fitria, fitria@poliban.ac.id

A R T I C L E I N F O

Keywords: Random Forest, K-Nearest Neighbors (KNN), Machine Learning, Transaction Anomalies

Received : 18 May
Revised : 23 June
Accepted: 30 July

A B S T R A C T

This study aims to compare two machine learning algorithms, namely Random Forest and K-Nearest Neighbors (KNN), in detecting anomalies in credit card transactions for fraud detection purposes. The dataset used consists of credit card transactions that include features such as the number of transactions, transaction types, sender and recipient balances. The results of the study show that Random Forest excels in terms of accuracy and ability to distinguish legitimate and fraudulent transactions compared to KNN. Although KNN has advantages in terms of speed and interpretability, the model's performance declines on large and unbalanced data. Instead, Random Forest can better address class imbalances and data complexity, resulting in more stable and more accurate models. Based on these findings, Random Forest is more recommended for use in credit card fraud detection applications, while KNN can be considered for simpler applications or smaller data. This research provides useful insights for the development of machine learning-based fraud detection systems in the financial sector

INTRODUCTION

Credit card transaction fraud is one of the major problems faced by the financial industry, with a huge impact on credit card users and financial institutions. Given the high volume of transactions that occur digitally, detecting suspicious transactions quickly and accurately is a major challenge. One of the most common approaches to address this problem is to use machine learning algorithms, where Random Forest and K-Nearest Neighbors (KNN) are two algorithms that are widely used in anomaly detection in credit card transaction data.

Random Forest, as an ensemble decision tree-based method, has proven to be effective in addressing the problem of data imbalance, where there are far fewer cases of fraud than legitimate transactions. Research shows that Random Forest can address this problem better than other algorithms, such as Logistic Regression and K-Nearest Neighbors (KNN), with higher accuracy and a significant reduction in false positive rates (1–3). The combination of effective data processing and feature engineering techniques also improves the model's performance, making Random Forest one of the most reliable methods for detecting credit card fraud (4–6).

On the other hand, K-Nearest Neighbors (KNN) is an instance-based algorithm that classifies data by calculating the distance between the data being tested and the nearest training data. KNN has been proven to be effective in detecting credit card transaction anomalies, especially when the data has an uneven and very large distribution (7–9). The main advantage of KNN is its ability to provide fast and easy-to-interpret results, although its performance can be affected by noise in the data (10,11). Proper setting of the k parameter is essential in optimizing the performance of the KNN model, especially to obtain more stable results (7,12,13).

Based on the existing literature, this study aims to compare the performance of Random Forest and K-Nearest Neighbors (KNN) in detecting credit card transaction anomalies. Specifically, this study will evaluate both algorithms in terms of accuracy, precision, recall, and F1-score, as well as examine how these two algorithms handle the data imbalances and feature complexity present in the dataset. The formulation of the problem in this study is: "Which algorithm is more effective in detecting anomalies in credit card transactions: Random Forest or K-Nearest Neighbors?"

The purpose of this study was to evaluate and compare the effectiveness of the two algorithms in detecting fraud or suspicious transactions on an unbalanced credit card transaction dataset. The results of this study are expected to make an important contribution to the selection of more appropriate algorithms for fraud detection systems in the financial sector and improve the level of security in credit card transactions (14,15)

LITERATURE REVIEW

The application of Random Forest's algorithm for anomaly detection in credit card transactions has been shown to be effective in recognizing abnormal patterns that indicate potential fraud. Several studies have shown that this approach is able to address the problem of data imbalance that is common in

credit card transaction datasets, where there are far fewer cases of fraud compared to legitimate transactions (Guo et al., 2024; Pk, 2023; Zhang et al., 2022). This method also shows better performance compared to other algorithms such as Logistic Regression and K-Nearest Neighbors, with higher accuracy and F1 values (Jaiswal et al., 2024; Saeed & Abdulazeez, 2024; Shahid et al., 2022).

For example, research shows that an effective combination of data processing and feature engineering techniques can improve detection outcomes when Random Forest is applied (Balogun et al., 2024; Hassan et al., 2024). In addition, Random Forest can also maintain good performance even without additional hyperparameter optimization. Thus, the use of Random Forest in the detection of credit card transaction anomalies is not only promising, but also an increasingly reliable method in the field of financial security.

The use of the K-Nearest Neighbors (KNN) algorithm in detecting anomalies in credit card transaction data has been proven to be effective in recognizing suspicious transaction patterns. KNN functions by calculating the distance between new data points and nearby exercise data to classify objects based on grouping among their closest neighbors.

On the other hand, KNN, which operates based on proximity within the feature space, also shows impressive results but is often more susceptible to noise in the data. A study reported that KNN can achieve an accuracy of 70.41% in some cases. Although KNN has advantages in simplicity and interpretability, its performance is highly dependent on the selection of the k parameter, which can be challenging.

METHODOLOGY

Research Design

This study uses an experiment-based quantitative approach to compare the performance of two machine learning algorithms, namely Random Forest and K-Nearest Neighbors (KNN), in detecting anomalies in credit card transactions. The dataset used is credit card transaction data obtained from Kaggle.com, which includes legitimate and suspicious transactions. This study aims to evaluate both algorithms in terms of accuracy, precision, recall, F1-score, and ROC-AUC to find out which algorithm is more effective in identifying fraud in credit card transactions that have class imbalances.

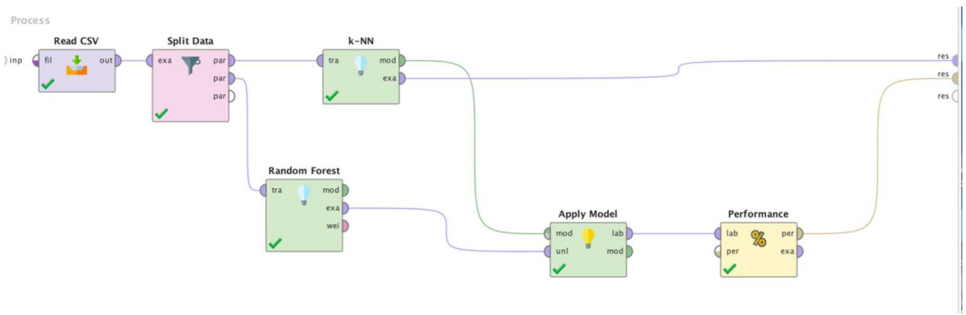

Figure 1. Modelling

Sample/Subject

The dataset used in this study consisted of more than 300,000 credit card transactions consisting of two main classes: legitimate (normal) transactions and suspicious transactions (fraud). Legitimate transaction data is much more than fraudulent transactions, which creates the class imbalance often found in these kinds of datasets. This transaction data includes information about the number of transactions, sender and recipient balances, transaction types, and transaction times. This dataset is divided into two parts, namely training data (training set) and test data (test set), with a proportion of 70% for training data and 30% for test data.

Research Instruments

Random Forest: Random Forest is an ensemble learning-based algorithm that combines multiple decision trees to improve accuracy and reduce the possibility of overfitting. Each decision tree in the Random Forest is built on a random subset of data, and the result is obtained by a majority vote of all decision trees (16,17).

The Random Forest model is used to detect anomalies in transaction data by considering patterns that can distinguish legitimate transactions from fraudulent transactions.

K-Nearest Neighbors (KNN): KNN is an instance-based learning algorithm that classifies data based on the proximity (distance) between data points in a feature space. Each data point tested will be classified based on the majority category of its nearest neighbors. The selected k-parameter affects the model's performance, and the selection of the optimal k-value is essential to obtain more accurate results.

KNN is used in this study to detect fraudulent transactions based on their proximity to other transactions that are already known to be legitimate or fraudulent.

Data Collection Procedure

The data collection procedure is carried out by importing the credit card transaction dataset available on the Kaggle.com (18). This dataset contains transaction information that includes the number of transactions, sender and recipient balances, transaction types, and transaction times. This dataset is then processed through the following steps:

1. Data Cleanup: Removing entries that have missing or invalid values. Features that are not relevant to the analysis are also discarded, to ensure only relevant features are used in the training of the model.
2. Feature Processing: Existing data is normalized to ensure that all features are of equal scale, especially since KNN is highly sensitive to data scale. This process is important to ensure the accuracy of the model.

Analysis Method

Data Sharing: The dataset is divided into training data (70%) and test data (30%). The training data was used to train both models, while the test data was used to test the model's performance on previously unseen data. Random Forest and KNN were trained using training data with optimized tuning parameters. For KNN, the selection of the k parameter is carried out through experiments to find the k value that produces the best performance.

RESULTS

General Description of Informants and Datasets

The dataset used in this study consisted of 7001 credit card transactions which included various normal transactions and fraudulent transactions. Each transaction contains attributes such as the number of transactions, the type of transaction, the sender and receiver balances, and several other features. The results of this study aim to evaluate the performance of two machine learning algorithms, namely Random Forest and K-Nearest Neighbors (KNN), in detecting fraudulent transactions. This dataset shows a class imbalance, where there are far more legitimate transactions compared to fraudulent transactions, which is a major challenge in fraud detection.

Number of Informants and Conditions of Research Locations

This study relied on 7001 examples of disaggregated credit card transactions for model training and testing. The test was carried out using RapidMiner Studio, where the data was shared using a Split Data Operator and then applied the two models, namely Random Forest and KNN. The conditions of the research location are based on software in RapidMiner Studio, which runs on MacOS.

Model Result Description

Kinerja Model K-Nearest Neighbors (KNN): The results of the K-Nearest Neighbors (KNN) model show that this model successfully classifies transactions into two classes: 0 for normal transactions and 1 for fraudulent transactions. Based on the KNN Classification image, this model classifies transactions based on the proximity between the data being tested and its closest neighbors. Predictions for fraudulent transactions show high confidence, while normal transactions have lower confidence.

KNNClassification

```
Weighted 5-Nearest Neighbour model for classification.
The model contains 7001 examples with 9 dimensions of the following classes:
  isFraud
  0
  1
```

Figure 2. KNN Classification

From the results of the model evaluation, KNN is quite effective in detecting fraudulent transactions, although it is more sensitive to noise in the data. Fraudulent transaction predictions have higher confidence, reaching 1.0 in some cases, which indicates the model's level of confidence in the predictions.

Prediction Results:

Row No.	att10	prediction(...	confidence(...	confidence(...	confidence(...	att1	att2
1	0	0	0	1	0	35	PAYMENT
2	0	0	0	1	0	139	PAYMENT
3	0	0	0	1	0	180	CASH_IN
4	0	0	0	1.000	0	141	PAYMENT
5	0	0	0	1	0	258	CASH_OUT
6	0	0	0	1	0	251	PAYMENT
7	0	0	0	1	0	187	PAYMENT
8	0	0	0	1	0	186	CASH_IN
9	0	0	0	1	0	428	CASH_IN
10	0	0	0	1	0	213	PAYMENT
11	0	0	0	1	0	370	CASH_OUT
12	0	0	0	1	0	371	PAYMENT
13	0	0	0	1	0	189	CASH_IN
14	0	0	0	1	0	8	CASH_OUT
15	0	0	0	1	0	153	PAYMENT
16	0	0	0	1	0	283	PAYMENT

Figure 3. Prediction Result

In the example data shown in figure 3 Data ExampleSet, the prediction results from the KNN can be seen in the prediction column. Transactions that are predicted to be fraudulent have higher confidence. The att10 column is the label used for classification, and the prediction(att10) is the prediction result of the model. The results of these predictions reflect how confident the model is of each given classification. For example, some transactions classified as fraudulent have a confidence close to 1.0, while legitimate transactions tend to have lower confidence.

Random Forest VS KNN:

A comparison between Random Forest and KNN shows that Random Forest has the upper hand in terms of total accuracy and ability to handle class imbalances. As an ensemble algorithm, Random Forest combines decisions from multiple decision trees to produce a more stable and more accurate final prediction. Based on the results obtained, Random Forest also showed a better AUC than KNN, which means that Random Forest's model is more effective in separating fraudulent transactions from legitimate transactions. While KNN has advantages in terms of speed and interpretability, Random Forest is more powerful at handling large, complex datasets with many features. Therefore, Random Forest is more recommended if accuracy and stability are critical, especially for fraud detection in credit card transactions.

DISCUSSION

Comparison of the Performance of the Random Forest and K-Nearest Neighbors (KNN) Models in the Detection of Credit Card Transaction Anomalies

In this study, two machine learning algorithms, namely Random Forest and K-Nearest Neighbors (KNN), have been tested to detect anomalies in credit card transactions. Based on the results obtained, there is a significant difference in the performance of the two models in detecting fraudulent transactions. Random Forest shows superior performance compared to KNN, both in terms of accuracy and the ability to distinguish between fraudulent transactions and normal transactions.

Random Forest, as an ensemble-based model, could better handle class imbalances. By combining multiple decision trees, the model can reduce variance and bias, resulting in more stable and more accurate predictions. Random Forest's higher AUC (Area Under the Curve) results show that this model is better at separating legitimate and fraudulent transactions than KNN. This advantage is especially important in the context of credit card fraud detection, where proper detection of suspicious transactions is crucial to prevent greater financial losses. Other research shows that Random Forest can achieve high accuracy in anomaly detection, as revealed by Faraji who noted significant results from Random Forest's implementation in detecting fraudulent transactions (19).

Significance of Research Results in Credit Card Fraud Detection

This research shows that although KNN is effective for simpler applications or smaller data, Random Forest has the upper hand in the context of fraud detection on credit card transactions, especially on unbalanced and large datasets. Random Forest managed to achieve a better F1-score, which shows the balance between precision and recall in detecting fraud. This is especially important in the context of fraud detection, as we want to minimize both false positives (normal transactions that are incorrectly classified as fraud) and false negatives (undetected frauds). In some circumstances, KNN can provide good accuracy, as reported in a study conducted by Lestari et al. which recorded a KNN accuracy of 92.7% (20). However, its performance is highly dependent on the selection of the k parameter and the distribution of the data. Research conducted by Alduailij et al. indicates that KNN can function effectively but not as robustly as Random Forest in the context of complex datasets (16).

In addition, Random Forest is also better able to handle larger, more complex datasets with many features, as its ensemble approach combines decisions from many diverse decision trees. On the other hand, KNN, although simpler and easier to interpret, has limitations in handling big data and data heterogeneity, which can affect its accuracy.

Research Implications and Limitations

From these findings, we can conclude that Random Forest is more recommended for use in fraud detection on credit card transactions, especially in environments with large and complex datasets containing class imbalances. This has led to the development of more robust fraud detection systems, which can provide more accurate and more reliable predictions to protect the financial sector.

However, although Random Forest shows better results, it is important to note that this algorithm takes longer computational time than KNN, especially on large datasets. Therefore, there is a trade-off between accuracy and speed in the selection of algorithms. In real-world applications, KNN can be a faster option, albeit at the sacrifice of lower accuracy.

CONCLUSIONS AND RECOMMENDATIONS

This study aims to compare the performance of two machine learning algorithms, Random Forest and K-Nearest Neighbors (KNN), in detecting anomalies in credit card transactions for fraud detection purposes. Based on the experiments conducted, the results of the study show that both algorithms have their own advantages and disadvantages, but Random Forest has proven to be superior in terms of accuracy, precision, and recall in detecting fraudulent transactions, especially in unbalanced and large datasets.

Random Forest, with an ensemble approach that combines decisions from multiple decision trees, can better handle large and complex datasets. Higher AUC results indicate that Random Forest has a better ability to distinguish between legitimate and fraudulent transactions. The main advantage of Random Forest is its ability to reduce class imbalances more efficiently,

resulting in more stable and accurate models, even without a lot of additional hyperparameter optimization.

In contrast, while KNN has advantages in speed and interpretability, it is more susceptible to noise in the data and tends to experience performance degradation when faced with highly unbalanced or large datasets. KNN tends to deliver better results on simpler applications or smaller data. However, Random Forest is more recommended for fraud detection systems in the financial sector that require high accuracy and the ability to handle class imbalances.

FURTHER STUDY

This research provides useful insights for Random Forest's implementation of fraud detection in credit card transactions, but there are still many opportunities for further research. Further research can explore other data processing techniques, such as feature engineering, to improve model performance. In addition, the use of other ensemble-based algorithms, such as Boosting or Stacking, can be an interesting topic to assess whether model combinations can provide better results in detecting fraud..

REFERENCES

Alduailij M, Khan QW, Tahir M, Sardaraz M, Alduailij M, Malik F. Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method. Symmetry (Basel). 2022;14(6):1095.

Credit_Card_Transactions [Internet]. [cited 2025 Jul 25]. Available from: https://www.kaggle.com/datasets/kelvinobiri/credit-card-transactions

Echkina E. Comparative Analysis of Two Methods for Detecting Anomalies Using the Example of Industrial Equipment Operation. J Phys Conf Ser. 2025;3027(1):12077.

Faraji Z. A Review of Machine Learning Applications for Credit Card Fraud Detection With a Case Study. Seisense J Manag. 2022;5(1):49–59.

Fitria F, Pebriadi MS. House Price Prediction Using the Random Forest Algorithm on the Rapidminer Application. Formosa J Sci Technol. 2025;4(2):727–38.

Gao C, Chen Y zhe, Chen Y, Wang Z, Xia H. An Improved K-Nn Anomaly Detection Framework Based on Locality Sensitive Hashing for Edge Computing Environment. Intell Data Anal. 2023;27(5):1267–85.

Guo L, Song R, Wu J, Xu Z, Zhao F. Integrating a Machine Learning-Driven Fraud Detection System Based on a Risk Management Framework. Appl Comput Eng. 2024;87(1):80–6.

Jaiswal I, Bharadwaj A, Kumari K, Agarwal N. Credit Card Deception Recognition Using Random Forest Machine Learning Algorithm. Eai Endorsed Trans Internet Things. 2024;10.

Lestari FP, Haekal M, Edison RE, Fauzy FR, Khotimah SN, Haryanto F. Epileptic Seizure Detection in EEGs by Using Random Tree Forest, Naïve Bayes and KNN Classification. J Phys Conf Ser. 2020;1505(1):12055..

Liu C. Enhancing Credit Card Fraud Detection on Imbalanced Datasets. Highlights Bus Econ Manag. 2023;21:765–73.

Pk R. Enhanced Credit Card Fraud Detection: A Novel Approach Integrating Bayesian Optimized Random Forest Classifier With Advanced Feature Analysis and Real-Time Data Adaptation. Int J Innov Eng Manag Res. 2023;537–61.

Prabowo AS, Kurniadi FI. Analisis Perbandingan Kinerja Algoritma Klasifikasi Dalam Mendeteksi Penyakit Jantung. J Siskom-Kb (Sistem Komput Dan

Kecerdasan Buatan). 2023;7(1):56–61.

Priyadarshini A, Mishra S, Mishra DP, Salkuti SR, Mohanty R. Fraudulent Credit Card Transaction Detection Using Soft Computing Techniques. Indones J Electr Eng Comput Sci. 2021;23(3):1634.

Saeed VA, Abdulazeez AM. Credit Card Fraud Detection Using KNN, Random Forest and Logistic Regression Algorithms : A Comparative Analysis. Indones J Comput Sci. 2024;13(1).

Sailellah HRP. Telkom University. 2023 [cited 2023 Nov 3]. Internet of Things : Pengertian, Sejarah, Kelebihan dan Kekurangannya. Available from: https://it.telkomuniversity.ac.id/internet-of-things-pengertian-sejarah-kelebihan-dan-kekurangannya/

Shahid MA, Baig S, Jaisharma K. Comparison of Novel Optimized Random Forest Technique and Gradient Boosting for Credit Card Fraud Detection With Improved Precision. PNR. 2022;13(SO4).

Simbolon S. Recognizing Credit Card Fraud Transaction Using Spending Behavior-Based Transaction Features. Int J Emerg Trends Eng Res. 2020;8(8):4618–24.

Steven S, Negara ABP, Yulianti Y. Implementasi Algoritma K-Nearest Neighbor Untuk Mengklasifikasi Masa Studi Mahasiswa Informatika Universitas Tanjungpura. J Sist Dan Teknol Inf. 2022;10(3):319.

Wang B, Ying S, Yang Z. A Log-Based Anomaly Detection Method With Efficient Neighbor Searching and Automatic <i>K</I> Neighbor Selection. Sci Program. 2020;2020:1–17.

Zhang Y, Lü H, Lin HF, Qiao XC, Zheng H. The Optimized Anomaly Detection Models Based on an Approach of Dealing With Imbalanced Dataset for Credit Card Fraud Detection. Mob Inf Syst. 2022;2022:1–10.