Fraud Detection In Credit Card Transactions Using HDBSCAN, UMAP And SMOTE Methods

Rudy Setiawan[1], Budi Tjahjono[2*], Gerry Firmansyah[3], Habibullah Akbar[4]

[1,2,3,4] Faculty of Computer Science, Universitas Esa Unggul, Indonesia
[*]Corresponding Author:
Email: budi.tjahjono@esaunggul.ac.id

Abstract.

Credit card abuse and fraud in credit card transactions pose a serious threat to financial companies and consumers. To overcome this problem, accurate and effective fraud detection is essential. In this study, we propose an approach that combines HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise), UMAP (Uniform Manifold Approximation and Projection), and SMOTE (Synthetic Minority Over-sampling Technique) methods to detect fraud in credit card transactions. The HDBSCAN method is used to group transactions based on their spatial density, allowing identification of suspicious groups of transactions. UMAP is used to reduce the dimension of transaction data, thus enabling better visualization and more efficient data analysis. In addition, we use SMOTE to overcome class imbalances, namely differences in the number of fraudulent and non-fraudulent transactions. In our experiments, we used. In this experiment, we used a dataset of credit card transactions that included both fraudulent and non-fraudulent transactions. The experimental results show that the proposed approach is able to detect fraud with high accuracy. The HDBSCAN method is able to effectively identify suspicious groups of transactions, while UMAP helps in better understanding and visualization of data. The use of SMOTE has successfully overcome class imbalances, resulting in more balanced fraud detection results between fraud and non-fraud. The results of this study show that the combination of HDBSCAN, UMAP, and SMOTE methods is effective in detecting fraud in credit card transactions. This approach can help financial companies identify suspicious transactions with high accuracy, reduce fraud losses, and improve the security of credit card transactions.

Keywords: *Fraud Detection, HDBSCAN, UMAP, SMOTE and Credit card transaction.*

I. INTRODUCTION

There are different types of fraud: insurance fraud, credit card fraud, statement fraud, securities fraud, and others. Of all, credit card fraud is the most common type. It is defined as unauthorized use of a credit card account. This occurs when the cardholder and card issuer are unaware that the card is being used by a third party[1]. In recent years, the volume of credit card transactions has increased dramatically due to the popularization of credit cards and the rapid development of e-services, including e-commerce, e-finance, and mobile payments.[2].Many companies that develop businesses use online media. The company provides facilities for online shopping for customers. Most customers use credit cards to buy things online. In this way, some customers can become thieves who have stolen someone's card to make online transactions. It is rated as credit card fraud that must be detected [3].The losses incurred from this fraud can reach billions of dollars. According to The Nilson Report, $28.65 billion was lost to payment card fraud worldwide in 2019, a 19.5% increase from $23.97 billion in 2017. Losses are projected to increase to 35.67 billion dollars in five years and 40.63 billion dollars in 10 years [4].According to the Federal Trade Commission (FTC), identity theft is the most common type of fraud reported by consumers, with credit card fraud accounting for 33% of all identity theft complaints in 2020[5]. The FTC further reports that credit card fraud has increased by 104% in the past decade, with 1.4 million fraud reports filed in 2020 alone.This surge in credit card fraud has led financial institutions to implement stronger fraud detection systems. Every day, new research is conducted by researchers in various fields. Many researchers in finance consider this issue to be challenging and important.

The use of machine learning is proposed by researchers to address this problem. The researchers trained machine learning algorithms with different types of data sets. This algorithm is proven to be helpful in detecting this kind of fraud [3].K-Means are commonly used for clustering because it is quick and easy to understand. However, there are some potential problems with this algorithm. First, the number of clusters, or partitions, needs to be provided as input to the algorithm. The number of clusters that may exist in a new dataset, however, is not always known [6].The local outlier factor (LOF) method was introduced by [7] which assigns the local affordability density to a data point and considers the density ratio between the point

and its nearest neighbor. LOF assigns an outlier score based on the local density of points. If the LOF outlier score is close to 1, the data point belongs to a very dense region that can be thought of as a cluster. If the LOF outlier score is significantly higher than 1, the point is more likely to be an outlier.In 2019 [8] proposed a new approach to credit card fraud detection using the DBSCAN clustering algorithm. The study shows that the proposed method outperforms traditional machine learning algorithms, such as Support Vector Machine (SVM) and Random Forest (RF), in terms of accuracy, precision, recall, and F1 score. The study also shows that the proposed method can detect previously unseen patterns of deception and reduce the number of false positives.The authors discuss the advantages of using the DBSCAN algorithm for credit card fraud detection, including its ability to detect local and global outliers and its toughness to noise. Further research could explore the effectiveness of combining DBSCAN with other outlier detection algorithms to detect credit card fraud.In 2021 [9] proposed a hybrid approach to credit card fraud detection that incorporates unsupervised and supervised learning techniques.

The unsupervised learning technique used is a clustering algorithm called Local Outlier Factor (LOF), which identifies outliers in the data set. The supervised learning technique used is the Random Forest algorithm, which is trained on non-outlier data to identify fraudulent transactions. The results showed that the proposed hybrid approach outperformed other methods in terms of accuracy, F1-score, and Area Under the Curve (AUC). The gap in this study is that it focused only on a limited set of unsupervised and supervised learning techniques. Future research may explore the effectiveness of unsupervised and other supervised learning techniques or the use of ensemble methods.Some data analysis methods that are commonly used to detect fraud in credit card transactions are by using clustering, outlier detection, and machine learning techniques. However, even though many data analysis techniques have been developed, there are still challenges in optimizing accuracy and speed in detecting fraud cases in credit card transactions.In this context, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is one of the clustering methods that is proven to be able to detect outliers in data with a high degree of accuracy. The combination of HDBSCAN with dimension reduction techniques such as Uniform Manifold Approximation and Projection (UMAP) and SMOTE undersampling techniques can be an alternative in overcoming challenges in detecting fraud cases in credit card transactions.In this study, fraud detection will be carried out on credit card transactions using a combination of HDBSCAN, UMAP, and SMOTE. The purpose of this study is to develop a data analysis model that can detect fraud cases in credit card transactions with a high level of accuracy, so as to help the financial industry in reducing fraud risks in credit card transactions and increasing customer trust in financial institutions.

II. METHODS

The initial stage of this research is by formulating research problems and determining research objectives, then continued with the second stage, namely a literature review related to the detection of outliers in credit card transactions to find research gaps and also state-of-the-art. The third stage is the establishment of an ML model for outlier detection in credit card transactions using secondary datasets obtained from Kaggle.com. The dataset used will be dimension-reduced using the UMAP method. to improve accuracy and performance in outlier detection on datasets. Dimensionality reduction aims to extract "latent" features from data that can represent original features in low-dimensional space. The technique is used to project data from high-dimensional space to low-dimensional space, maintaining the quality of the presentation. PD can reduce the computational time of the model and make it easier to visualize outlier detection results [10]. Uniform Manifold Approximation and Projection (UMAP) is another nonlinear dimensionality reduction technique that assumes data is uniformly distributed on Riemannian manifolds, locally constant Riemannian metrics and locally connected manifolds [11].By using dimensionality reduction techniques for outlier detection, it can reduce the high dimensionality of data so that it identifies outliers more easily. This can be very useful for identifying anomalies in datasets with high-dimensional features or multiple variables [12].Resampling aims to equalize the number of instances per class to either reduce majority class instances, known as under-sampling, or increase minority class instances, known as over-sampling.

The under-sampling technique reduces majority instances by randomly eliminating majority class instances [13]. With the advantage that it can increase run time and storage problems. However, it can eliminate important data. The remaining data may be biased samples and cannot provide accuracy for class distribution [14]. . Various under-sampling methods have been proposed, and the most commonly used are random under-sampling and Tomek link (T-link). The Synthetic Minority Oversampling Technique (SMOTE) is an improved oversampling technique developed by [15]. SMOTE is based on k's nearest neighbor to produce new synthetic sampling in feature spaces based on a certain percentage for minority classes. SMOTE can generate new synthetic data based on existing minority class data without replicating it to overcome the challenge of overfitting. The algorithm that works in SMOTE will first take the value of the difference between the vector of the feature in the minority class and the nearest neighbor value of the minority class and multiply that value by a random number between 0 and 1. Next, the calculation results are added with the feature vector so that a new vector value is obtained.The HDBSCAN clustering algorithm is a density-based algorithm. Unlike KMeans, it does not require that every data point be assigned to the cluster, as it identifies the density of the cluster. Points not assigned to the cluster are considered outliers, or noise [16]. Algorithms that can effectively find different groups in a dataset and identify outliers are valuable techniques.

The HDBSCAN algorithm process consists of the following stages:

1. First, the algorithm will build a k-nearest neighbor graph for each data point in the dataset. This graph will be used to calculate the data density at each point in the dataset.
2. Next, the algorithm will perform hierarchical clustering or hierarchical clustering on the nearest neighbor graph using an approach called condensed tree. Hierarchical clustering aims to combine points in the graph of nearby neighbors that have the same density or close to.
3. After that, the algorithm will select the minimum cluster size value and cut the condensed tree at a certain level. This is done to obtain different clusters and remove noise in the dataset.
4. Finally, the algorithm will group each data point in the dataset into identified clusters.

III. RESULT AND DISCUSSION

A. Data Collecting and preprocessing

At this stage, data collection was carried out from various dataset sources, one of which was through the Kaggle public dataset site, where the author obtained a dataset in the form of credit card transaction data. This dataset contains transactions, occurring within two days, made in September 2013 by cardholders in Europe. The dataset contains 284,807 transactions of which 492 transactions are fraudulent and the rest are genuine. Considering the numbers, we can see that this dataset is very unbalanced, where only 0.173% of transactions are labeled as fraudulent. The dataset consists of 31 columns consisting of a transaction time column, a column for the number of transactions made, a class column that indicates fraud or non-fraud transactions, and columns V1-V28 which are confidential transaction data. Figure 1 show the distribution of fraud and non fraud data.

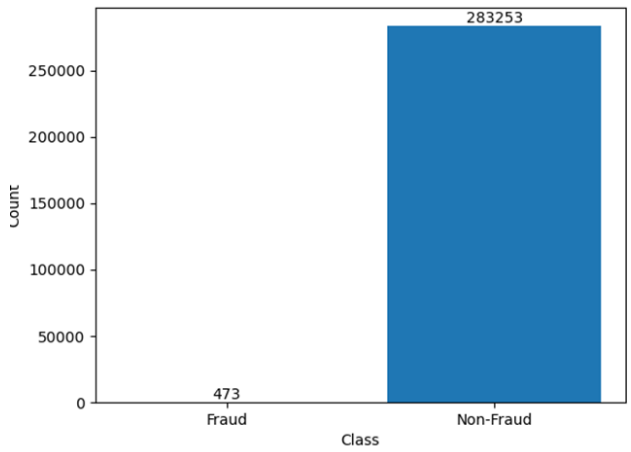

Fig 1. Comparison of the amount of fraud and non-fraud data

Next is preprocessing the dataset, in this phase, we will first scale the column consisting of Time and Amount. Time and sum should be scaled as other columns. This is done so that the dataset becomes easier to understand. On the other hand, we also need to create a subsample of the dataframe to have the same number of fraud and non-fraud cases.Due to the imbalance between data fraud and non-fraud, the application of SMOTE will be carried out to overcome the data imbalance. From the results of using SMOTE, the results of conditioning the number of normal data 2000 and the amount of data fraud 400 were obtained. Table 1 show data sampling with SMOTE.

Table 1. Sampling with SMOTE

Fraud	Non Fraud	Total
400	2000	6000

Furthermore, the dataset will be reduced using the UMAP method, this is done to speed up the learning process. V1-V28 feature data will be combined using UMAP method into 2 new columns UMAP_1 and UMAP_2. Furthermore, V1-V28 features will be removed from the dataset. Table 2 shows example values in the dataset, Table 3 shows the results of conditioning with UMAP.

Table 2. Examples of existing values in a dataset

Time	Amount	V1	V28	Class
0.0	149.62	-0.072781	-0.021053	0
0.1	2.69	0.266151	0.014724	0
0.2	378.66	-1.340163	-0.059752	0

Table 3. Conditioning results with UMAP

Scalled_time	Scalled_amount	UMAP_1	UMAP_2
0.220275	-0.189245	13.532091	1.729030
2.416910	-0.532457	13.332249	2.220625
-0.097344	0.550884	8.697934	5.735172

B. Establishment of Fraud Detection Model Using HDBSCAN

HDBSCAN is used for the formation of fraud detection models, HDBSCAN has 2 hyperparameters that can be changed, namely "min_cluster_size" and "min_samples [17]. In this study, the author made settings on these hyper-parameters to find out the best parameter values. For the use of min cluster size the author uses random numbers between 10 – 250, and for min samples the author uses random numbers between 5 – 130. Furthermore, the results of the iteration will be evaluated using the AUC-Score, precission, recall and F1-Score evaluation methods.Figure 2 show the comparison result using different hyper parameter.

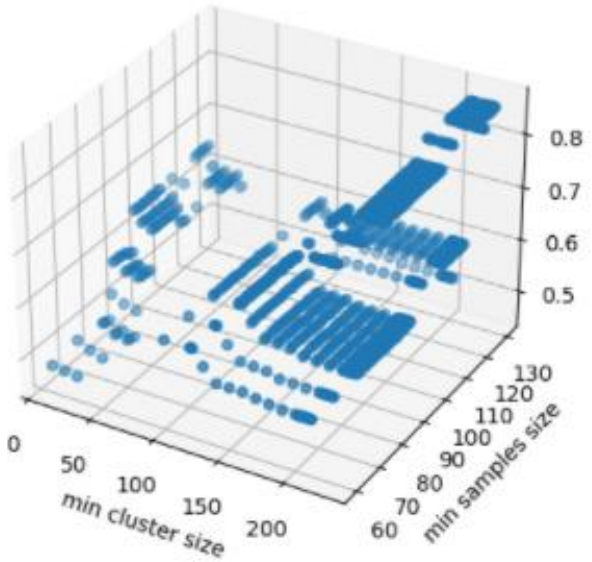

Fig 2. AUC-Score comparison results using different hyper parameters

C. Model Evaluation

In this process, the author uses a combination of parameters that have been tested before, getting a good combination of parameters for the model to be trained. To evaluate the model that has been built, the author calculates the value of precision score, recall score and f1 score [18] [19] [20] with the following formula.

$$Accuracy = \frac{number\ of correctly\ classified\ classes}{total\ number\ of\ classes} \times 100\%$$

$$Precission = \frac{True\ Positive}{True\ Positive + False\ Positive} \times 100\%$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \times 100\%$$

$$F1\ Score = 2 \times \frac{Precission \times Recall}{Precission + Recall}$$

From the results of the experiments that have been carried out, the evaluation results are as follows:

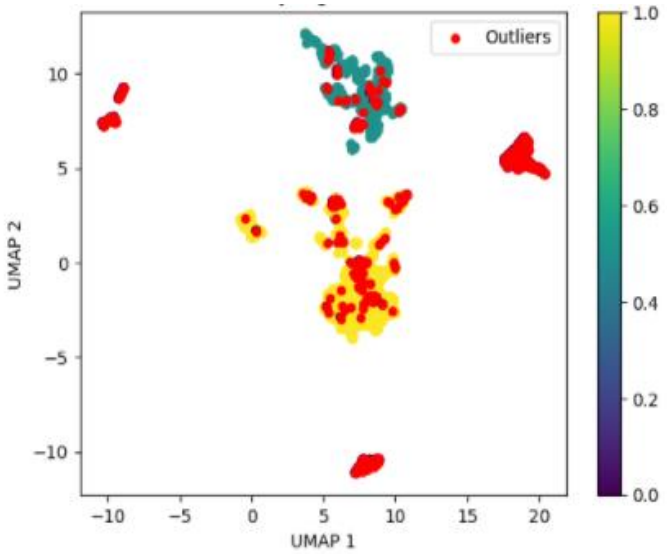

Fig 3. Number of outliers found

The final result of the AUC value is formed into an ROC curve which can be seen in Figure 4.

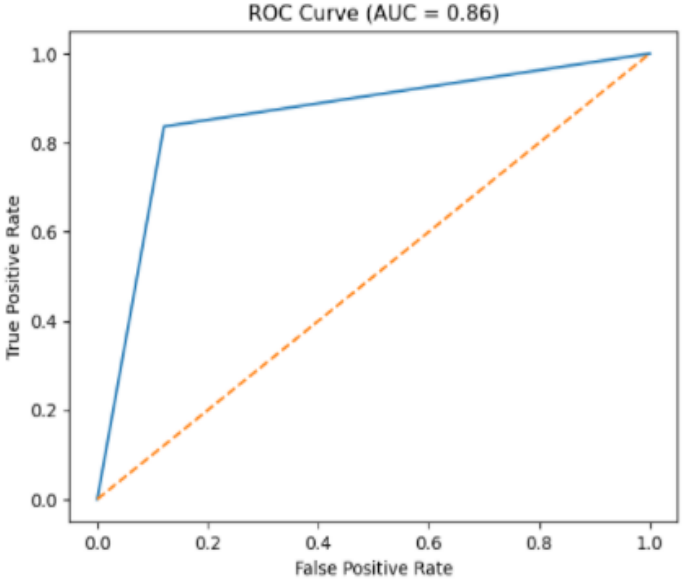

Fig 4. ROC Curve

Overall, the experimental results are shown in Table 2. The table shows that combined techniques namely HDBSCAN, UMAP and SMOTE have good performance, evaluation using evaluation metrics namely precision, recall, F1-Score and AUC-Score.

Tabel 4. Model Evaluation Metrics

Total Outlier	Precision	Recall	F1 Score	AUC score
601	54%	84%	65%	86%

IV. CONCLUSION

Based on the results of the study above, it can be concluded that the right combination of hyper parameters can produce a high AUC Score value in the model built, the highest AUC score in this study is 86% with min cluster size parameters 211 and min samples 110. Thus this model can detect fraud in credit card transactions with good results.

V. ACKNOWLEDGMENTS

The author is grateful to many parties who have helped and supported this research.

REFERENCES

[1] S. A. Z. T. Y. H. R. H. M. S. &. Z. H. Makki, "Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M. S., & Zeineddine, H. (2019). An experimental study with imbalanced classification approaches for credit card fraud detection. IEEE Access, 7, 93010-93022.," 2019.

[2] X. H. Y. X. W. &. W. Q. Zhang, "HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture," Information Sciences, vol. 557, pp. 302-316, 2021.

[3] P. C. &. G. S. T. Cynthia, "An outlier detection approach on credit card fraud detection using machine learning: a comparative analysis on supervised and unsupervised learning," Intelligence in Big Data, Springer, https://doi.org/10.1, 2021.

[4] Simamora, R. N. H., & Elviani, S. (2022). Carbon emission disclosure in Indonesia: Viewed from the aspect of board of directors, managerial ownership, and audit committee. ***Journal of Contemporary Accounting***, 1-9.

[5] B. M. Y. O. B. &. M. Q. Itri, "Composition of Feature Selection Methods And Oversampling Techniques For Banking Fraud Detection With Artifical Intelligence.," ***International Journal of Engineering***, vol. 11, pp. 216-226, 2021.

[6] "Federal Trade Commission. (2021). Consumer Sentinel Network Data Book 2020. Retrieved from https://www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-2020/consumer_sentinel_network_data_book_2020.pdf".

[7] G. Stewart dan M. Al-Khassaweneh, "An implementation of the HDBSCAN* clustering algorithm.," Applied Sciences, 12(5), 2405., 2022.

[8] Tarigan, N. M. R., Syahputra, R. A., & Yudha, T. K. (2022). The Analysis of Quality of Work Life and Work Achievement in Department of Agriculture Simalungun Regency. SIASAT, 7(1), 55-70.

[9] M. M. Breunig, H. P. Kriegel, R. T. Ng dan J. Sander, "LOF: identifying density-based local outliers," dalam In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 93-104)., 2000, May.

[10] GEMINTANG, "Automatic Credit Card Fraud Detection System Using Dbscan Outlier Detection," etd.repository.ugm.ac.id, http://etd.repository.ugm.ac.id/penelitian/detail/176098, 2019.

[11] L. B. Y. A. C. O. K. Y. O. F. &. B. G. Carcillo, "Combining unsupervised and supervised learning in credit card fraud detection.," Information sciences, vol. 557, pp. 317-331, 2021.

[12] Sarkum, S., Syamsuri, A. R., & Supriadi, S. (2020). The role of multi-actor engagement. ***Journal of Open Innovation***: Technology, Market, and Complexity, 6(4), 176.

[13] O. Vlasovets, "Unsupervised anomaly detection in merchant vessel data.," 2020.

[14] L. McInnes, J. Healy dan J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction.," arXiv preprint arXiv:1802.03426., 2018.

[15] L. Weijler, F. Kowarsch, M. Wödlinger, M. Reiter, M. Maurer-Granofszky, A. Schumich dan M. N. Dworzak, "UMAP based anomaly detection for minimal residual disease quantification within acute myeloid leukemia.," Cancers, 14(4), 898., 2022.

[16] K. M. I. M. S. S. F. M. A. M. J. P. M. H. S. M. I. H. A. S. &. R. O. Hasib, "A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem.," ***Journal of Computer Science***, , pp. 16(11), 1546-1, 2020.

[17] Supriadi, ., Dalimunthe, R. F., Lumbanraja, P., & Tarmizi, H. B. (2021). The Antecedent Of Educational Staff Contextual Performance In Medan City Private Universities. Archives of Business Research, 9(2), 316–338. https://doi.org/10.14738/abr.92.9817

[18] P. Fergus, D. Huang dan Hamdan, "Prediction of intrapartum hypoxia from cardiotocography data using machine learning," Applied Computing in Medicine and Health—Emerging Topics in Computer Science and Applied Computing, pp. Volume 1, pp. 125–146, 2016.

[19] N. Chawla, K. Bowyer, L. Hall dan W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," ***J. Artif. Intell.***, pp. 16, 321–357, 2002.

[20] R. J. Campello, M. D. GB dan S. J., "Density-based clustering based on hierarchical density estimates.," Lecture Notes in Computer Science, 7819, 160-172., 2013.

[21] Sri, R., Mahdi, F., Julkarnain, J., Kurnia, H. N. T., & Habibie, A. (2022). Intellectual capital and islamic corporate social responsibility on the financial performance of sharia commercial banks in Indonesia. In E3S Web of Conferences (Vol. 339, p. 05003). EDP Sciences.

[22] S. M. N. B. M. K. A. &. M. A. Mishra, "An Evaluative Measure of Clustering Methods Incorporating Hyperparameter Sensitivity," Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 7, pp. 7788, (2022, June).

[23] S. C. &. Z. S. Tana, "Binary search of the optimal cut-point value in ROC analysis using the F1 score.," 2019.

[24] D. F. S. A. &. M. G. A. Kakhki, "Evaluating machine learning performance in predicting injury severity in agribusiness industries.," Safety science, 117, 257-262., 2019.

[25] K. M. &. L. J. Rashid, "Times-series data augmentation and deep learning for construction equipment activity recognition.," Advanced Engineering Informatics, 42, 100944., 2019.