Indonesian Journal of Electrical Engineering and Computer Science
Vol. 37, No. 2, February 2025, pp. 1014~1022
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v37.i2.pp1014-1022

Enhancing data cleaning process on accounting data for fraud detection

Mohamad Affendi Abdul Malek, Kamarularifin Abd Jalil
College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia

Article Info

Article history:

Received Jul 24, 2024
Revised Sep 17, 2024
Accepted Sep 29, 2024

Keywords:

Accounting data
Accounting data cleaning
Data cleaning
Dataset
Fraud detection
Fraudulence data

ABSTRACT

Data cleaning is a crucial step in fraud detection as it involves identifying and correcting any inaccuracies or inconsistencies in the data. This can help to ensure that the data being used for fraud detection is reliable and accurate, which in turn can improve the effectiveness of fraud detection algorithms. Due to the overwhelming amount of data, data cleaning specific for fraud detection is a very important activity for the auditor to find the appropriate information. Therefore, a new accounting data cleaning for fraud detection is needed. In this paper, an enhancement of the process of fraud detection by accounting auditors through the implementation of accounting data cleaning technique is proposed. The proposed technique was embedded in a prototype system called accounting data cleaning for fraud detection (ADCFD). Through experiment, the performance of the proposed technique through ADCF is compared with those obtained from the IDEA system, using the same dataset. The results show that the proposed enhanced technique through ADCFD system performed better than the IDEA system.

Corresponding Author:

Kamarularifin Abd Jalil
College of Computing, Informatic and Mathematics, Universiti Teknologi MARA
40450 Shah Alam, Selangor, Malaysia
Email: kamar446@uitm.edu.my

1. INTRODUCTION

Data cleaning or also known as data scrubbing is the process of correcting and amending or removing data in a dataset that is incomplete, incorrect, improperly formatted, or duplicated or otherwise problematic (‘dirty’) data and records [1]. An organization in a data-intensive field like banking, insurance, retailing, telecommunications, or transportation might use a data cleaning tool to systematically examine data for flaws by using rules, algorithms, and look-up tables. Other researchers quote that data cleaning refers to standardizing data from different sources (unstructured file format) to a common format so that data can be better utilized and analyzed [2]-[5]. The goal of data cleaning is to prepare the data for use in modelling or analyzing data whereas this study focused on fraud detection analysis.

Poor data quality is a fact of life for most organizations and can have serious implications on their effectiveness. An example critical application domain is accounting, where incorrect information about payments in a payment account may lead to read flags up to fraudulence data. A recent approach for repairing dirty data is to use data quality rules in the form of database constraints to identify tuples with errors and inconsistencies and then use these rules to derive updates to these tuples. Most of the existing data repair approaches focus on providing fully automated solutions using different heuristics to select updates that would introduce minimal changes to the data, which could be risky especially for critical data. To guarantee that the best desired quality updates are applied to the database, users (domain experts) should be involved to confirm updates [6].

In this paper, a new data cleaning process is proposed for accounting data in detecting fraud. With effective cleaning, all data sets should be consistent and free from any errors that could be problematic during later use or analysis. The paper starts by reviewing the fraud detection techniques and data cleaning process. The paper then explains the methodology used to develop the proposed data cleaning process before explaining it in detail afterward. This is followed by presenting the results from the experimental tests conducted to compare between the proposed data cleaning process with a conventional data cleaning software and followed by the conclusion.

2. FRAUD DETECTION

The Oxford Dictionary defines fraud as “the crime of cheating somebody in order to get money or goods illegally”. This definition encapsulates the core of fraud and encompasses the numerous forms and types of deception. On the other hand, it does not provide much direction for considering the requirements of a fraud detection system because it does not explicitly identify the nature and features of fraud [7].

2.1. Fraud detection techniques and procedure

Fraud detection techniques and procedures are methods used to identify and prevent fraudulent activities in financial transactions or businesses. These techniques involve analyzing patterns and data to identify any anomalies or suspicious behavior, such as unusual transaction amounts or account activity [8]. Fraud detection procedures typically include data collection, analysis, and reporting, as well as internal and external controls to prevent fraudulent activities. Common fraud detection techniques include rule-based systems, machine learning (ML) algorithms, and artificial intelligence tools [7], [9]-[11]. Additionally, various fraud prevention measures, such as two-factor authentication and identity verification, can also help to reduce the risk of fraudulent activities.

In their study, Krishna and Praveenchandar [12] assert that conducting a comparative examination of credit card fraud detection using logistic regression (LR) with random forest (RF) leads to an enhanced accuracy in predicting fraudulent activities. According to Nobakht *et al.* [13], synthetic data creation and data-driven modelling are used in the development of a fraud detection system for gasoline bunkering. Van Belle *et al.* [14] assert that Inductive Graph Representation learning is employed for the purpose of detecting fraudulent activities within credit card transaction networks. In their study, Honghao *et al.* [15] discuss the optimization of weighted extreme learning machines for unbalanced classification, specifically in the context of credit card fraud detection. Baesens *et al.* [7] assert that data engineering is crucial for the purpose of fraud detection.

2.2. Application software for fraud detection

Application software for fraud detection can assist businesses and organizations to identify and prevent fraudulent activities. Such software employs advanced analytics and ML algorithms to analyze large datasets, detect suspicious transactions or behavior, and generate alerts in real-time. There are various types of fraud detection software available in the market, including credit card fraud detection, insurance fraud detection, and bank fraud detection. According to Deng *et al.* [16], they propose a data mining approach to detect transaction fraud. In their study, Liang *et al.* [17] assert that the utilization of synthetic data generation and data-driven modelling is crucial in the creation of a fraud detection system specifically designed for fuel bunkering.

3. ACCOUNTING DATA CLEANING

Accounting data cleaning is the process of identifying and correcting errors, inconsistencies, and inaccuracies in financial records to maintain the accuracy of the overall financial statement. Accounting transactions are complex and prone to errors during manual entry [18]. Therefore, using automated tools and software can help you identify data discrepancies, such as duplicate entries or missing values, and make necessary corrections in real-time.

3.1. Data cleaning techniques and procedure

Data cleaning techniques are used to identify and remove or correct errors, inconsistencies, and other issues in a dataset to improve its quality and accuracy. Some common procedures include data normalization, duplicate detection and removal, handling missing values, outlier detection and treatment, and data transformation. The specific techniques used depend on the nature of the data and the goals of the analysis. It is an important step in data preprocessing that ensures the data is ready for analysis.

Chu *et al.* [19] presented a taxonomy of the data cleaning literature, emphasizing the growing interest in qualitative data cleaning strategies that employ constraints, rules, or patterns to identify problems.

Zhang *et al.* [20] presented a data cleaning model utilizing the Ensemble Learning technique for the original condition monitoring dataset of power equipment. Subsequently, a distance discriminant technique is employed to identify anomalous data by comparing the predicted findings with the measured values. Following this, the missing data is subsequently imputed.

In their study of the data cleaning process, Ridzuan and Wan Zainon [21] emphasized the issues associated with expanding data in the context of big data, as well as the various solutions available for data cleaning. In their study, Kumar and Khosla [22] emphasized the examination and depiction of unorganized data pertaining to air pollution. They also performed a survey to explore techniques for eliminating undesirable, contaminated, and unclear data.

3.2. Application software for data cleaning

Application software for data cleaning helps to organize, manipulate, and verify data by removing redundancies, correcting errors, and updating records. Some of the commonly used data cleaning application software include Microsoft Excel, OpenRefine, Trimble, and Talend. These programs can help to improve the accuracy and completeness of the data, which in turn can enhance the efficiency of data analysis. It is also important to note that data cleaning should be done on an ongoing basis to ensure that the data remains accurate and up to date.

Tee *et al.* [23] assert that system E-clean is founded upon a data cleaning architecture designed for patient data. In their study, Wang [24] focused on categorizing time series data cleaning approaches, examining the most advanced methods for different types of data cleaning, analyzing the tools and systems used for data cleaning, discussing assessment criteria employed in both research and industry, and proposing future directions for time series data cleaning. In their study, Tee *et al.* [23] emphasized that the E-clean system utilizes the extract, transform, and load (ETL) model as its primary process model. This model serves as a clear framework for implementing the system. In addition, parsing techniques are also employed for the detection of corrupt data. The selected approach for attribute matching is regular expression. The k-nearest neighbor (KNN) algorithm is chosen from the available data cleaning algorithms.

3.3. Data quality assurance

data quality assurance (DQA) is the process of ensuring that the data being used in an organization is accurate, complete, consistent, reliable, timely, and secure. It involves a set of policies, procedures, and tests that are designed to identify and correct data issues before they become problems. DQA is important because poor quality data can lead to incorrect decisions, inefficiencies, and even financial losses. Effective DQA helps organizations to maintain the integrity of their data and to make better use of their information assets.

Zhang *et al.* [20] presented a data cleaning model that utilizes the Ensemble Learning technique (namely, random forest (RF)) to process the original condition monitoring dataset of power equipment. Subsequently, a distance discriminant technique is employed to identify anomalous data by comparing the predicted findings with the measured values. Following this, the missing data is subsequently imputed. Zou [25] emphasized that data cleaning of raw data is a crucial step in the big data analysis and application process.

In their study, Bezmenov [26] emphasized various methods for identifying outliers in data cleaning, such as statistical, probabilistic, and ML techniques. Nevertheless, Zou [25] exclusively concentrated on broad categories of datasets. However, the study conducted by Bezmenov [26] exclusively concentrated on the process of data cleaning specifically for the Romanian home energy sector. Bi *et al.* [27] solely concentrated on the widespread use of information collecting and transmission technology in the electricity system.

The prevailing method for detecting application fraud often commences with ETL processes. A data cleaning intervention is incorporated into the algorithm by utilizing parsing techniques, regular expressions, KNN, and RF. An evaluation approach can assess the efficacy of fraud detection in electronic accounting data following the data cleaning process. The post-data cleaning evaluation approach should focus on assessing the efficacy of fraud detection by analyzing the integrity of identified fraudulent accounting data. Performing accounting data cleaning would enhance the accuracy of identifying false accounting data, hence improving the effectiveness of fraud detection analysis. Implementing advanced fraud detection algorithms and data evaluation methods significantly improved the effectiveness of detecting fraudulent accounting data. Based on the research conducted, it was found that it is necessary to enhance the existing model for fraud detection procedure. In this paper, a conceptual framework for cleaning accounting data is proposed as shown Figure 1.

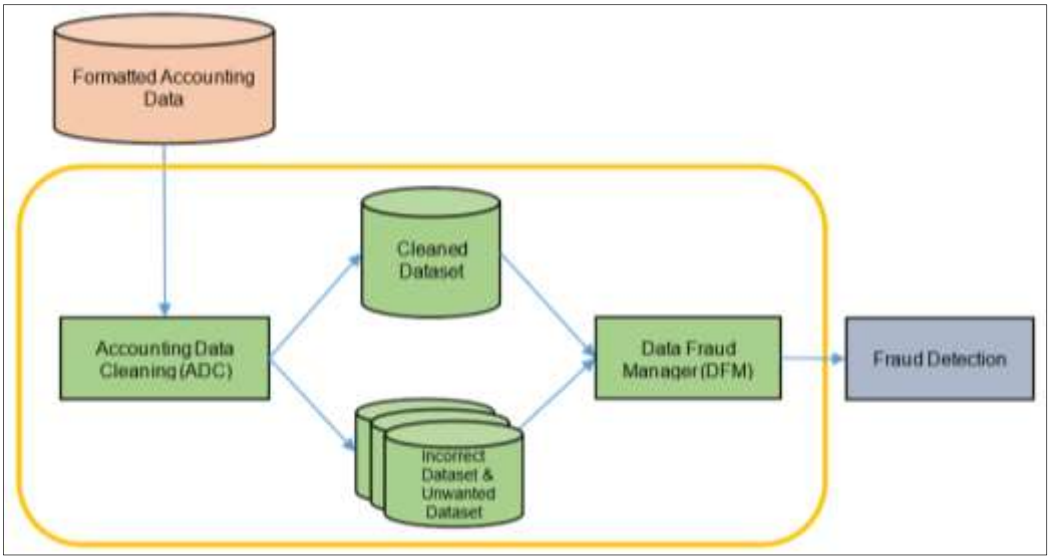
Formatted Accounting Data
Cleaned Dataset
Accounting Data Cleaning (ADC)
Data Fraud Manager (DFM)
Fraud Detection
Incorrect Dataset & Unwanted Dataset

Figure 1. Proposed conceptual framework based on accounting data cleaning
(Adapted from Ibrahim & Mustafa Kamal, 2018)

4. METHOD

An experimental study method is specifically applied using Account Payable data, utilizing different processes of data cleaning. This suggests a deliberate effort to explore and compare various data cleaning techniques in the context of fraud detection. The study aims to employ common fraud detection techniques, showcasing a comprehensive approach to assessing the proposed model's impact. Interestingly, the use of Account Payable data suggests a real-world and practical application of the model within the financial domain. This choice of data type also implies a recognition of the significance of cleaning accounting data in the context of detecting fraudulent activities. The method used is depict in Figure 2.

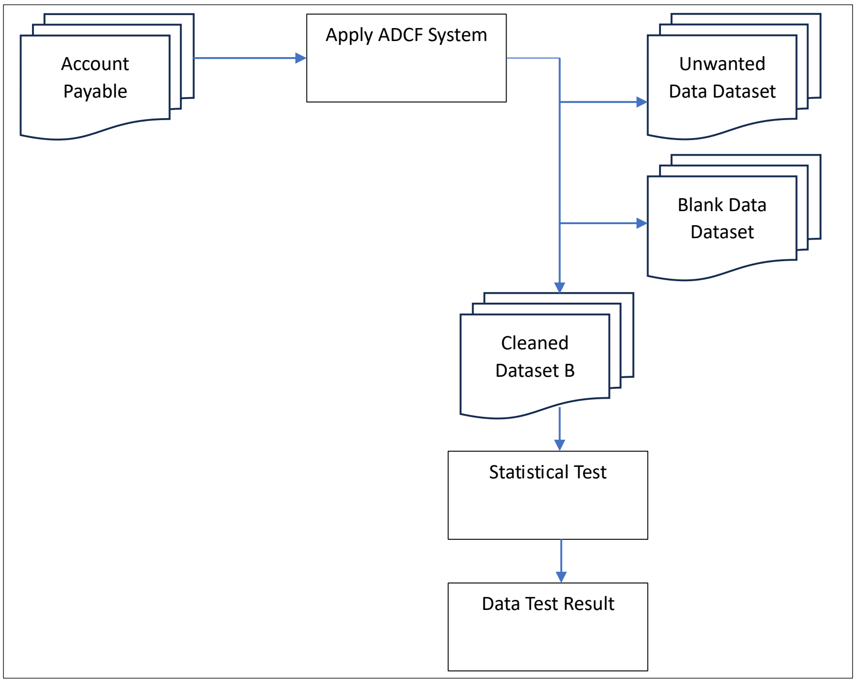

Figure 2. ADCF design

4.1. Measures

Several data cleaning methods have an impact on the fraud detector. This study would not address all of these topics due to the vast scope of this field. This study specifically examines two primary aspects of data cleaning: blank data and unwanted date values as shown in Table 1.

The analysis based on this mapping is stated in the table above. However, it would start with a data extraction process. This is one of the processes in ETL. To develop a data cleaning model for fraud detector, the process follows the strategy base on the mapping table using accounting raw data collection of account payable.

Table 1. Accounting data cleaning versus fraud detection method mapping table

	Data cleaning	Fraud detection method
1.	Rule violation (Blank data)	
2.	Pattern violation (Unwanted date value)	Aging analysis

4.2. Data analysis

A standard notation system was used:

- X: indicates a group’s exposure to an experimental variable or event, the consequences of which would be measured.
- O: indicates a measurement or observation made on an instrument.

Xs and Os in each row are applied to the same specific functions. Xs and Os in the same column, or placed vertically relative to each other, are simultaneous. The left-to-right dimension represents the temporal order of procedures in the experiment (sometimes represented with an arrow). To test O1 and O2 researcher used statistical analysis which is t-test. A t-test is a statistical method used to determine whether there is a significant difference between the means of two independent samples. It is commonly used when the population standard deviations are unknown. The test statistic, represented by t, compares the differences between the sample means to the variability within the samples. The null hypothesis of the test is that the two-sample means are equal, while the alternate hypothesis is that at least one sample mean is different from the other. The T-test can be used to test both one-tailed and two-tailed hypotheses.

5. THE PROPOSED DATA CLEANING PROCESS

Data cleaning is a crucial step in fraud detection as it involves identifying and correcting any inaccuracies or inconsistencies in the data. This can help to ensure that the data being used for fraud detection is reliable and accurate, which in turn can improve the effectiveness of fraud detection algorithms. Some common data cleaning techniques used in fraud detection include data normalization, data standardization, data imputation, and outlier detection. By using these techniques, analysts can improve the quality of their data and better identify patterns and trends related to fraudulent activity.

The current fraud detection algorithm such as foundation models, clustering, ML (decision tree (DT), KNN, LR, RF), convolutional neural network (CNN), anomaly detection, inductive graph representation learning, synthetic data generation, and modelling techniques enhance accuracy in the accounting fraud detection. Implementing an advanced fraud detection system that utilizes fraud detection algorithms and data evaluation methods significantly improved the accuracy and efficiency of identifying fraudulent accounting data. The performance of the fraudulent data cleaning and detection system should be assessed by measuring its accuracy, recall, precision, matthews correlation coefficient, and the area under the curve. This evaluation should be conducted using ML algorithms combined with the adaptive boosting (AdaBoost) technique. Suspicious transaction patterns should be identified, annotated on a graph, and analysis. The system improves the detection of fraudulent data by utilizing fraud detection techniques and ML algorithms, namely the AdaBoost methodology.

In this paper, the enhancement to the process for cleaning data in accounting is proposed as shown into Figure 3. According to the proposed mechanism, in order to improve a fraud detection system, we should begin by implementing an ETL process. Next, perform data cleaning on any identified fraudulent accounting data. Then, employ parsing techniques and regular expressions to classify the quality of the fraudulent accounting data using various algorithms. Finally, assess the effectiveness of the detected fraudulent data. In order to test the enhanced data cleaning process, a prototype system to design the proposed accounting data cleaning for fraud detection prototype system called ADCF was developed.

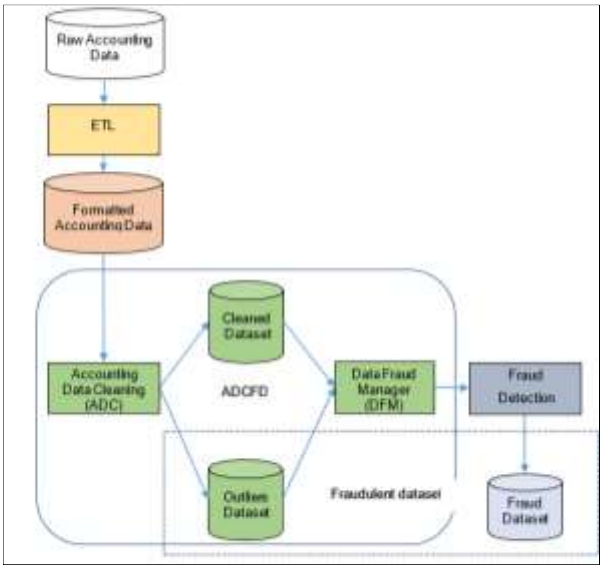

Figure 3. Proposed process of accounting data cleaning

6. RESULTS

In order to evaluate the proposed enhancement of the data cleaning process, an experiment has been conducted. The experiment is actually comparing the performance of the existing data cleaning process using the IDEA® caseware with the proposed enhanced cleaning process using ADCF. The metrics performance for the experiment includes an examination of the prototype's design, functionalities, and how well it aligns with the intended goals of cleaning accounting data for fraud detection purposes. This measurement is achieved by comparing the outcomes of the ADCFD system with those obtained from the IDEA system, using the same dataset. This comparative analysis provides insights into the effectiveness and performance of the ADCFD system in relation to an established benchmark.

The experiment used Microsoft Excel to perform data cleaning tasks. Microsoft Excel offers a built-in functionality that enables users to eliminate empty rows and columns simultaneously. The filter function is frequently employed in order to extract particular data from a sizable dataset. Nevertheless, it can also serve the purpose of eliminating blank cells. Table 2 shown the output of IDEA system after data cleaning using Microsoft Excel whereas Table 3 shows the output of cleaning the data using the ADCF proposed system which employed the enhanced process for cleaning data.

Table 2. Aging result in IDEA system after data cleaning using Microsoft Excel

Int (Days)	#Records	(%) Records	Net value	(%) Net value
0	2,623	41.06	99,53,107	48.75
30	249	3.9	8,79,430	4.31
60	262	4.1	9,72,991	4.77
90	248	3.88	9,07,701	4.45
120	242	3.79	8,66,605	4.24
150	306	4.79	10,72,271	5.25
180	276	4.32	9,21,395	4.51
180+	2,182	34.16	48,42,048	23.72
ERR	0	0	0	0
Totals:	6,388	100	2,04,15,548	100

Table 3. Aging result in IDEA system after data cleaning using ADCF prototype system

Int (Days)	#Records	(%) Records	Net value	(%) Net value
0	2,579	40.77	97,71,161	48.44
30	248	3.92	8,73,041	4.33
60	260	4.11	9,64,092	4.78
90	248	3.92	9,07,701	4.5
120	242	3.83	8,66,605	4.3
150	306	4.84	10,72,271	5.32
180	274	4.33	9,16,667	4.54
180+	2,169	34.29	48,01,171	23.8
ERR	0	0	0	0
Totals:	6,326	100	2,01,72,709	100

The dataset used served as the baseline for auditors to conduct aging analysis, identifying discrepancies between expected and actual results. However, with the integration of the ADCFD model, auditors now yield two additional sets of data that are considered as potentially fraudulent datasets. Consult Table 4 for a detailed examination of the outcomes produced by the ADCFD model and the IDEA® software. This comparative analysis aims to underscore the enhanced capability of the ADCFD model in unveiling fraudulent patterns beyond the scope of conventional data cleaning and existing software tools, providing auditors with a more comprehensive understanding of potentially fraudulent activities within the datasets under scrutiny.

Table 4. Data frequency of ADCF and IDEA® Software

Software	Data type	Frequency
ADCFD prototype system	Blank data	3
	Unwanted data	62
IDEA system	Cleaned data	6326

In this paper, the analysis technique use to compare the performance of these two methods is the t-test, a widely used statistical method for assessing the significance of differences between means in two groups. The t-test is applied to scrutinize the datasets based on the output list of the IDEA system, as presented in Table 2 and Table 3. These tables likely contain essential information and metrics representing the characteristics of the datasets both pre and post ADCFD application. By leveraging the t-test on the IDEA system's output, the study aims to discern whether there are statistically significant variations in the datasets before and after the ADCFD process. This quantitative approach provides a robust means of evaluating the impact and efficacy of the ADCFD model in enhancing the detection of potentially fraudulent activities within the examined datasets. The researcher used t-test function in Microsoft Excel and the output of t is: t = 0.097217.

$$t = \frac{(\sum D)/_{N}}{\sqrt{\frac{\sum D^2 - \left(\frac{(\sum D)^2}{N}\right)}{(N-1)(N)}}}$$

Using the statistical analysis, particularly the application of the independent samples t-test to compare two groups, where µ1 and µ2 represent the means of the two groups being compared. The null hypothesis posits that these means are equal, implying an equivalence in the context of the groups under consideration. Additionally, the discussion touches upon the paired t-test, where the null hypothesis asserts that the difference between the two tests is equal to zero (H0: µd = 0), indicating no significant difference between the paired observations.

Moving to the results of the analysis, the t-value is computed and found to be 0.097217. The interpretation of this t-value involves comparing it to the critical table value at a significance level of 0.05. The fact that the computed t-value exceeds the critical table value implies that the null hypothesis, which suggests there is no difference between the means, must be rejected. In other words, the statistical evidence indicates that there are indeed differences in the dataset of the ADCFD prototype system before and after its application. The rejection of the null hypothesis suggests that the ADCFD system has a discernible impact on the dataset, affirming differences in the dataset before and after the application of the ADCFD process.

7. CONCLUSION

The proposed enhancement of the data cleaning process which is applied to the ADCFD model plays a pivotal role in refining data for the purpose of enhancing fraud detection capabilities. Its key feature lies in the efficient handling of data selection and management, particularly with regard to blank and unwanted data. Unlike outright removal, the proposed process employs a strategic approach whereby identified blank and unwanted data are transferred to another dataset. This transfer is contingent upon specific selection criteria, including the recognition of blank data and the identification of unwanted data. By adopting this selective data management strategy, the system aims to provide auditors with a nuanced and comprehensive dataset during the fraud detection process. The intended result is an improvement in the efficiency of data cleaning for fraud detection purposes.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Teknologi MARA for funding this research.

REFERENCES

[1] E. Rahm and H. Do, "Data cleaning: problems and current approaches," *IEEE Data Engineering Bulletin*, vol. 23, no. 4, pp. 3–13, 2000.

[2] N. S. Chauhan, "Data cleaning: challenges and existing solutions," *Interantional Journal of Scientific Research in Engineering and Management*, vol. 08, no. 04, pp. 1–5, Apr. 2024, doi: 10.55041/ijsrem30377.

[3] S. Guha, F. A. Khan, J. Stoyanovich, and S. Schelter, "Automated data cleaning can hurt fairness in machine learning-based decision making," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–12, 2024, doi: 10.1109/TKDE.2024.3365524.

[4] P. O. Côté, A. Nikanjam, N. Ahmed, D. Humeniuk, and F. Khomh, "Data cleaning and machine learning: a systematic literature review," *Automated Software Engineering*, vol. 31, no. 2, p. 54, Nov. 2024, doi: 10.1007/s10515-024-00453-w.

[5] Q. Yuan *et al.*, "ULDC: unsupervised learning-based data cleaning for malicious traffic with high noise," *Computer Journal*, vol. 67, no. 3, pp. 976–987, Apr. 2024, doi: 10.1093/comjnl/bxad036.

[6] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas, "Guided data repair," *Proceedings of the VLDB Endowment*, vol. 4, no. 5, pp. 279–289, 2011, doi: 10.14778/1952376.1952378.

[7] B. Baesens, S. Höppner, and T. Verdonck, "Data engineering for fraud detection," *Decision Support Systems*, vol. 150, p. 113492, Nov. 2021, doi: 10.1016/j.dss.2021.113492.

[8] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: a review of anomaly detection techniques and recent advances," *Expert Systems with Applications*, vol. 193, p. 116429, May 2022, doi: 10.1016/j.eswa.2021.116429.

[9] O. Odeyemi, C. V. Ibeh, N. Z. Mhlongo, O. F. Asuzu, K. F. Awonuga, and F. O. Olatoye, "Forensic accounting and fraud detection: a review of techniques in the digital age," *Finance & Accounting Research Journal*, vol. 6, no. 2, pp. 202–214, Feb. 2024, doi: 10.51594/farj.v6i2.788.

[10] A. K. Lin, "The AI revolution in financial services: emerging methods for fraud detection and prevention," *Jurnal Galaksi*, vol. 1, no. 1, pp. 43–51, May 2024, doi: 10.70103/galaksi.v1i1.5.

[11] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, "Enhancing Credit Card Fraud Detection: An Ensemble Machine Learning Approach," *Big Data and Cognitive Computing*, vol. 8, no. 1, p. 6, Jan. 2024, doi: 10.3390/bdcc8010006.

[12] M. V. Krishna and J. Praveenchandar, "Comparative analysis of credit card fraud detection using logistic regression with random forest towards an increase in accuracy of prediction," in *International Conference on Edge Computing and Applications, ICECAA 2022 - Proceedings*, Oct. 2022, pp. 1097–1101, doi: 10.1109/ICECAA55415.2022.9936488.

[13] S. B. B. Nobakht, Y. Liang, G. Lindsay, "Data-driven modelling for condition-based monitoring and flow regime prediction in flow systems," in *Proceedings of the 39th International North Sea Flow Measurement Workshop*, 2021, p. 13.

[14] R. Van Belle, C. Van Damme, H. Tytgat, and J. De Weerdt, "Inductive graph representation learning for fraud detection," *Expert Systems with Applications*, vol. 193, p. 116463, May 2022, doi: 10.1016/j.eswa.2021.116463.

[15] J. Honghao, Y. Po, and Y. Tianyu, "The influence of adolescents' romantic relationship on individual development: Evidence from China," *International Journal of Chinese Education*, vol. 10, no. 3, Sep. 2021, doi: 10.1177/22125868211070036.

[16] W. Deng, Z. Huang, J. Zhang, and J. Xu, "A data mining based system for transaction fraud detection," in *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, Jan. 2021, pp. 542–545, doi: 10.1109/ICCECE51280.2021.9342376.

[17] Y. Liang, B. Nobakht, and G. Lindsay, "The application of synthetic data generation and data-driven modelling in the development of a fraud detection system for fuel bunkering," *Measurement: Sensors*, vol. 18, 2021, doi: 10.1016/j.measen.2021.100225.

[18] M. J. Kranacher, R. A. (Dick) Riley, and J. T. Wells, "Forensic accounting and fraud examination," *European University Institute*, no. 2, pp. 2–5, 2012, [Online]. Available: https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32016R0679&from=PT%0Ahttp://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52012PC0011:pt:NOT.

[19] X. Chu, I. F. Ilyas, S. Krishnan, and J. Wang, "Data cleaning: Overview and emerging challenges," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Jun. 2016, vol. 26-June-2016, pp. 2201–2206, doi: 10.1145/2882903.2912574.

[20] S. Zhang, W. Yao, P. Sun, and Y. Zhang, "A condition monitoring data cleaning method for power equipment based on correlation analysis and ensemble learning," in *7th IEEE International Conference on High Voltage Engineering and Application, ICHVE 2020 - Proceedings*, Sep. 2020, pp. 1–4, doi: 10.1109/ICHVE49031.2020.9279409.

[21] F. Ridzuan and W. M. N. Wan Zainon, "A review on data cleansing methods for big data," *Procedia Computer Science*, vol. 161, pp. 731–738, 2019, doi: 10.1016/j.procs.2019.11.177.

[22] V. Kumar and C. Khosla, "Data cleaning-a thorough analysis and survey on unstructured data," *Proceedings of the 8th International Conference Confluence 2018 on Cloud Computing, Data Science and Engineering, Confluence 2018*, pp. 305–309, 2018, doi: 10.1109/CONFLUENCE.2018.8442950.

[23] L. K. Tee, C. Chee, H. H. Mohamed, and O. S. Lee, "E-clean: a data cleaning framework for patient data," *Proceedings - 1st International Conference on Informatics and Computational Intelligence, ICI 2011*, pp. 63–68, 2011, doi: 10.1109/ICI.2011.21.

[24] X. Wang and C. Wang, "Time series data cleaning: a survey," *IEEE Access*, vol. 8, pp. 1866–1881, 2020, doi: 10.1109/ACCESS.2019.2962152.

[25] F. Zou, "Research on data cleaning in big data environment," in *Proceedings - 2022 International Conference on Cloud Computing, Big Data and Internet of Things, 3CBIT 2022*, Oct. 2022, pp. 145–148, doi: 10.1109/3CBIT57391.2022.00037.

[26] I. V. Bezmenov, "Method of cleaning outliers from measurement data: search for the optimal solution with the minimum number of rejected measured data," *Measurement Techniques*, vol. 66, no. 1, pp. 14–23, Apr. 2023, doi: 10.1007/s11018-023-02184-y.

[27] Y. Bi, X. Zhao, Y. Zhou, L. Lao, and S. Jiang, "Factors associated with the depression among people with disabilities: a cross-sectional study in Chinese communities of Shanghai," *Medicine (United States)*, vol. 99, no. 47, p. E23331, Nov. 2020, doi: 10.1097/MD.0000000000023331.

BIOGRAPHIES OF AUTHORS

Mohamad Affendi Abdul Malek is a Ph.D. candidate at Universiti Teknologi MARA, Shah Alam, Selangor, Malaysia. He is currently teaching at Universiti Teknologi MARA, Raub, Pahang. He can be contacted at email: maffendi@uitm.edu.my.

Kamarularifin Abd Jalil received his first degree in Computer Science from Universiti Teknologi MARA in 1995. He then further his studies to a Master level in 1997 from the Coventry University in the field of Information Technology for Manufacture. In 2002, he further his studies to a Ph.D. level at University of Strathclyde in the field of Mobile Communication. He has been working as a lecturer with University Teknologi MARA since 1997. His research interests include computer networks and network security. He can be contacted at email: kamar446@uitm.edu.my.