Indonesian Journal of Electrical Engineering and Computer Science
Vol. 32, No. 2, November 2023, pp. 838~844
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v32.i2.pp838-844

The nexus of corruption and non-performing loan: machine learning approach

Suraya Masrom[1], Rahayu Abdul Rahman[2], Nurul Huda Mohd Shukri[3], Naimah Ahmad Yahya[2], Mohd Zulfikri Abd Rashid[2], Nor Balkish Zakaria[4]
[1]Computing Sciences Studies, College of Computing, Informatics and Media, Universiti Teknologi MARA, Tapah, Malaysia
[2]Faculty of Accountancy, Universiti Teknologi MARA, Tapah, Malaysia
[3]Institute of Professional Studies, Universiti Poly-Tech MARA, Kuala Lumpur, Malaysia
[4]Accounting Research Institute, Universiti Teknologi MARA, Shah Alam, Malaysia

Article Info

Article history:

Received Mar 22, 2023
Revised Jul 26, 2023
Accepted Aug 4, 2023

Keywords:

Banking institution
Corruption
Machine learning
Malaysia
Non-performing loan

ABSTRACT

Banking institutions around the world are facing a serious problem with non-performing loans (NPLs), which can jeopardize their financial stability and hinder their ability to issue new loans. The issue of NPLs has been linked to corruption, which has emerged as one of the contributing factors. Given the scarcity of research on the use of machine learning (ML) techniques to examine the relationship between corruption and NPLs, this paper provides an empirical evaluation of various ML algorithms for predicting NPLs. Besides ML performance comparisons, this paper presents the analysis of ML features importance to justify the effect of corruption factor in the different ML algorithms for predicting NPLs. The results indicated that most of the tested ML algorithms present good ability in the prediction models at accuracy percentages above 70% but corruption index has contributed very minimal effect to the ML performances. The most outperformed ML algorithm in the different proposed settings is random forest. The framework of this research is highly reproducible to be extended with a more in-depth analysis, particularly on problems of NPL as well as on the ML algorithms.

Corresponding Author:

Rahayu Abdul Rahman
Faculty of Accountancy, Universiti Teknologi MARA
Tapah Campus, 34500 Tapah, Perak, Malaysia
Email: rahay916@uitm.edu.my

1. INTRODUCTION

Non-performing loans (NPLs) have posed a substantial challenge for banks across the world [1]. According to Orlando and Pelosi [2], NPLs are the loans that have either defaulted or are on the brink of default, indicating that the borrower may not repay the loan according to its initial terms. Previous research has shown that a number of macroeconomic factors, including as the global financial crisis [3], [4], the COVID-19 pandemic [5], as well as unemployment and inflation rates [3], [6]–[9], contribute to NPLs. Recently, some academics [10]–[12] have emphasized the effect that corruption has on NPLs. Desta [13] defined corruption as an unethical or illegal use of power for personal gain, often involving practices like embezzlement, fraud, bribery, and nepotism [13]. It erodes public trust in government, public and private institutions [14] and can result in negative outcomes for individuals and society, such as reduced economic growth and social inequality.

The banking sector, in particular, is vulnerable to corruption which eventually erodes the public's confidence in the financial system. It's because corruption leads to an unfair distribution of opportunities and resources, particularly during the credit allocation process, which raises the risk of loan default and increases the amount of NPLs. Indeed, Son *et al.* [15] contend that corruption in the allocation of credit can lead to loans being given to borrowers who are not qualified or deserving of them because corrupt bank employees or officials may take part in or overlook dishonest loan practices like falsifying loan documents, or insider trading.

The NPLs have a major and detrimental impact on bank performance in a number of ways, including decreased bank profitability, capital, and lending capacity [16]. It's because having a lot of NPLs can result in larger loan loss provisions, which can lower a bank's net revenue and profitability. This situation will also result in a decline in the bank's capacity to make new loans as the bank may need to set aside more capital to cover loan losses. As a result, the bank's reputation will suffer, which will lower confidence among shareholders, staff, and customers. In Malaysia, NPLs have been a substantial issue for the banking sector. The NPL ratio in Malaysia has been gradually increasing since the late 1990s, with some fluctuations [17]. At the same time, the country continues to have difficulty combating corruption on all levels. Malaysia was rated 61st out of 180 nations in the 2021 corruption perception index (CPI), with a score of 52 out of 100 [18], despite the government's efforts to combat corruption in the nation. The CPI index measures the public's opinion of official corruption in a given country. The lower the CPI score, the greater the public view of the chance of Malaysian officials being involved in corruption.

Thus, it is becoming crucial for banking institutions, stakeholders, and regulators to anticipate NPLs with corruption. Machine learning (ML) has recently emerged as a critical predictive mechanism in a number of fields, including in education [19], [20], fraud detection [21] and medical [22]. Also, a number of studies [23]–[25] have demonstrated that ML can produce findings that are more accurate when predicting NPLs. For instance, Bellotti *et al.* [23] uses a private database from a European debt collection firm to compare the results of a diverse range of regression approaches and ML algorithms for forecasting recovery rates on non-performing loans. The results show that rule-based algorithms such as cubist, boosted trees and random forests perform significantly better than other approaches. Meanwhile, Serengil *et al.* [24] apply various ML algorithms to develop NPL prediction models on a customer portfolio dataset in a private bank in Turkey. Using a large dataset, 181,276 samples, different performance metrics; precision, recall, F1 score, imbalance accuracy (IAM), specificity has been performed. The findings reveal that light gradient boosting machine (LightGBM) gave the best results among the logistic regression, support vector machine (SVM), random forest (RF), bagging classifier, eXtreme gradient boosting (XGBoost) and long short-term memory (LSTM) for the dataset.

Moreover, Elnaggar *et al.* [25] used an Egyptian credit dataset with 112,907 occurrences and 17 characteristics to predict non-performing loans. For training and testing purposes on the dataset, classification techniques such as logistics regression (LR), k-nearest neighbors (KNN), SVM, decision tree (DT), and meta-classifier have been used in this study. The findings demonstrate that the ensemble method's accuracy outperforms all others.

Notwithstanding the widespread use of ML in NPLs, the inclusion of corruption is extremely scarce in the literature. Recent studies on corruption-NPLs mostly relied on conventional essential approaches mainly multiple linear regression [10]–[12]. Therefore, the objective of this study is to extend the state-of-the-art of corruption by focusing on ML NPLs prediction models. The existence of advanced techniques of ML can be deployed immensely and to start with a few of them is highly crucial. Besides identifying the ML performances through prediction accuracy and time efficiency, another important question that needs to be answered is how the corruption aspect contributed to the performances of different ML models. This issue has not been broadly discussed in the existing research reports.

2. METHOD

2.1. Sample of data

This study uses 351 of bank year observations listed on the Malaysian stock exchange as a sample dataset. Similar to prior research [11]–[13], this study uses a corruption index to measure corruption. Particularly, this study uses the following formula to measure the degree of annual corruption of the country: CI=100-CI index scores by transparency international. There are 10 features that derived from two main categories; bank specific characteristics (bank size (size), bank liquidity (liquidity), management efficiency (MEf), bank profitability (ROAA and ROAE)), credit risk and Islamic banking status and macroeconomics (corruption (CI), gross domestic product growth rate (GDPG) and annual inflation rate (INF)). The dependent variable is NPL that is defined as impaired loans to gross loans ratio. The correlation coefficient from Pearson correlation test of each feature to the dependent variable was identified as listed in Table 1. Corruption index has the highest correlation, which is expected as the most important feature importance in the ML prediction models for predicting the NPL.

Table 1. Features weights of correlations to NPL

Features/Variables	Meaning	Measurement method	Correlation
Corruption index	Corruption index is the perceived levels of public-sector corruption in a country	CI=100-CI index scores by transparency international.	0.464
GDPG	GDP growth rate	GDP growth rate.	0.004
INF	Annual inflation rate	Inflation rate.	0.203
ROAA	Return on average assets	ROAA is defined as net income over the average assets.	0.177
ROAE	Return on equity	ROAE is defined as net income over the average equity.	0.157
Size	Bank size	The bank size is defined as the logarithm of total assets.	0.231
Liquidity	Liquidity ratio	The liquidity ratio equals the loan divided by customer's deposits.	0.208
MEf	MEf	MEf is measured by dividing the operating expenses by the operating income generated, signaling bank efficiency in managing their operation by reducing costs and increasing profit.	0.200
IB	Classification of Islamic bank and non-Islamic bank	Islamic banks are assigned as 1, while others equal to 0.	0.079
CR	Credit risk	CR is defined as the ratio of doubtful debts to total loans.	0.456

2.2. The machine learning algorithms

Based on auto model RapidMiner preliminary study, four out of six suggested ML algorithms were selected to be compared namely DT, random forest, gradient booster and SVM. The rest of two algorithms namely generalized linear model and deep learning performed less than these four algorithms at mean absolute error (MAE) above 0.4 and relative error above 50%. All the four algorithms have a few hyper-parameters, which are optimized by the auto model RapidMiner. Figure 1 shows that the best maximal depth for DT was 4 to achieve and the lowest relative error rate can be achieved is 43.5%. Furthermore, for random forest (refer Figure 2), the best error rate was 41.2% with maximal depth 7, number of trees 100. The lowest error for gradient boosted trees was 42.8 with setting in Figure 3 while for SVM was 41.8%, which depicted in Figure 4. The hyper-parameters settings were used by researchers to develop manual ML models in RapidMiner to observe the performances based on different split ratios.

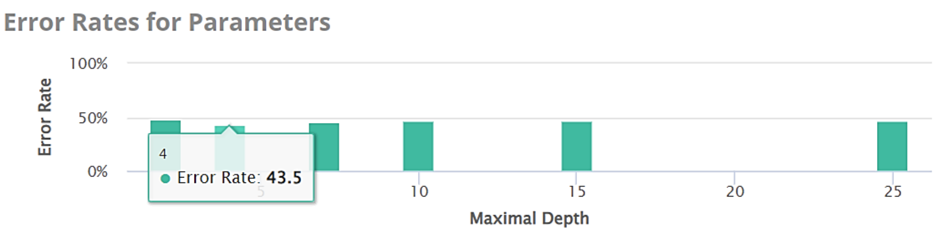

Figure 1. The best configuration for decision tree

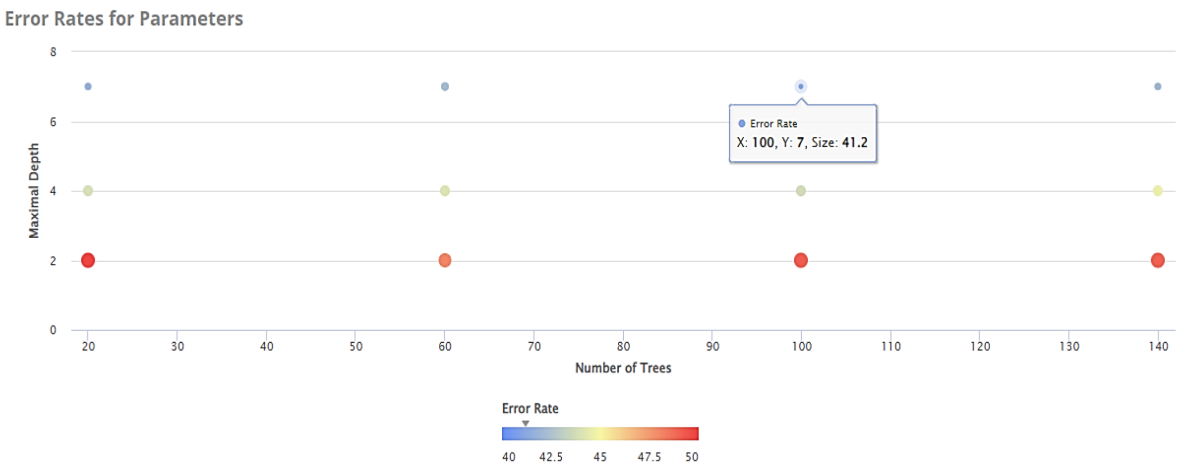

Figure 2. The best configuration for random forest

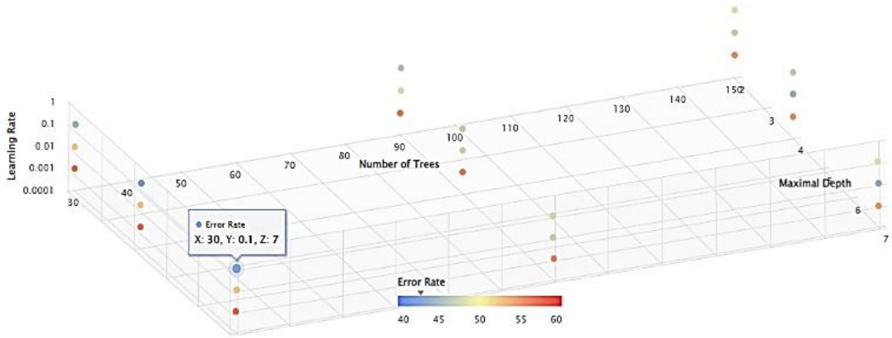

Figure 3. The best configuration for gradient boosted trees

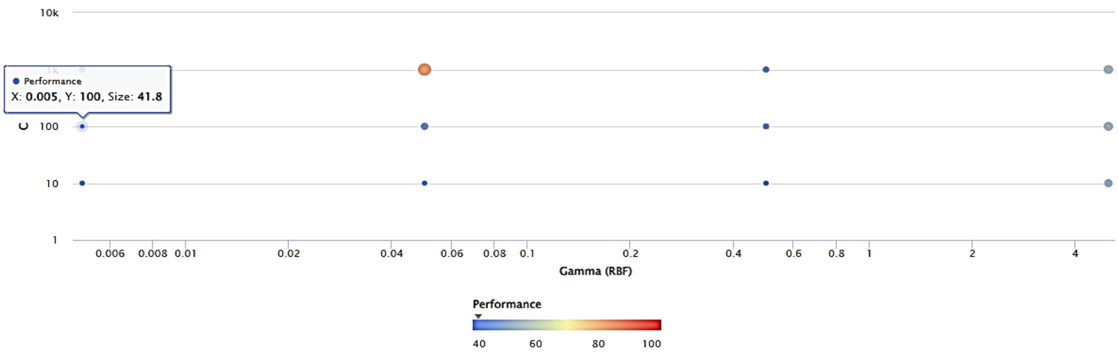

Figure 4. The best configuration for SVM

2.3. Training approach and evaluation techniques

Separating the collected dataset into training and testing is an important step for evaluating the ML algorithms. From the 351 records, 211 were used for training and the rest 140 were used as the hold-out samples for testing the ML prediction ability. Thus, the split ratio between training and testing was 60:40. Root mean square error (RMSE) and MAE were used to measure the model accuracy while R squared (R^) can be used as an additional metric to indicate the total features importance in the NPL prediction models.

Based on the preliminary findings from auto model RapidMiner, manual processes were developed as depicted in Figure 5. The process begins with reading the NPL dataset, identifying the role (DV), split data according to different groups of split ratios (60:40, 70:30, 80:20), training the ML algorithms, testing (apply model), and comparing the performances.

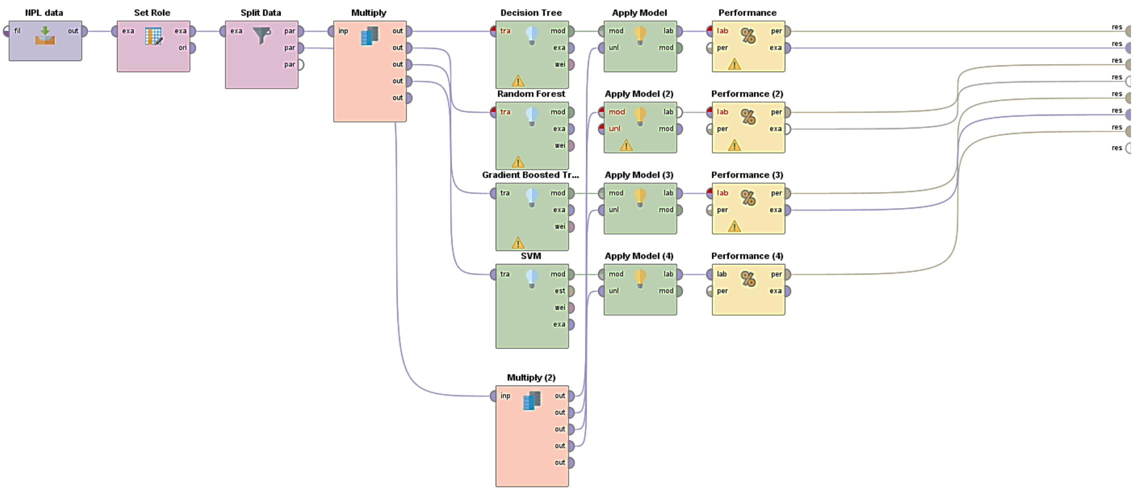

Figure 5. Developing machine learning models in RapidMiner

3. RESULTS AND DISCUSSION

The results were divided into two. Firstly, the results of the performances of the ML in predicting NPL are given in Table 2. Secondly, how important is the corruption index in comparison with other features in each of the ML algorithms will be presented.

Regardless of ML algorithms, the results in Table 3 show a significant performance improvement from the split ratio 60:40 to 70:30 but consistent achievement at 80:20. Slight improvement by SVM from split ratio 70:30 to 80:20. The most outstanding ML algorithm is random forest with the highest R^ and lowest RMSE and MAE at all the split ratios. Considered split ratio 70:30 as the best setting for most of the algorithms, this setting was chosen to be furtherly compared in accordance to the features importance as given in Table 3. In Table 3 present the most feature importance is MEf for DT, random forest, and gradient boosted trees. The corruption index is the most important feature only in SVM. Some features have no correlation coefficient weights to the NPL namely IB in random forest, ROAE, and CR in SVM.

Table 2. The performance results

Algorithm	R^	RMSE	MAE
Split ratio 80:20			
Decision tree	0.72	3.43	3.29
Random forest	0.77	3.26	2.55
Gradient boosted tree	0.65	4.31	3.01
Support vector machine	0.75	4.21	2.55
Split ratio 70:30			
Decision tree	0.72	3.43	3.29
Random forest	0.77	3.28	2.57
Gradient boosted tree	0.65	4.37	3.01
Support vector machine	0.73	4.33	2.75
Split ratio 60:40			
Decision tree	0.70	4.86	3.48
Random forest	0.77	3.46	2.61
Gradient boosted tree	0.64	4.77	3.12
Support vector machine	0.70	4.64	2.88

Table 3. The weights of correlations of each NPL feature

Feature	Decision tree	Random forest	Gradient boosted trees	Support vector machine
Corruption index (CI)	0.075	0.192	0.061	0.300
GDPG	0.069	0.079	0.051	0.004
INF	0.017	0.005	0.019	0.028
ROAA	0.010	0.014	0.010	0.009
ROAE	0.038	0.000	0.113	0.000
Size	0.014	0.004	0.093	0.162
Liquidity	0.039	0.006	0.120	0.001
MEf	0.420	0.210	0.384	0.068
IB	0.002	0.000	0.005	0.057
CR	0.032	0.005	0.087	0.000

4. CONCLUSION

This paper presents significant findings of research concerned with NPLs and corruption indexes as two important issues in banking institutions. Although corruption has a higher relationship to NPL generally beyond ML prediction models, the findings from this research found minimal effects to most of the selected ML algorithms. Therefore, further research is important to get insight into the issue with different ML learning algorithms. The DT, random forest, and gradient boosted trees are categorized as tree-based features selection mechanism, thus focusing on linear based ML might be the potential research to be conducted in the future.

ACKNOWLEDGEMENTS

The authors would like to thanks the Universiti Poly-Tech Malaysia, grant code: 100-TNCPI/PRI 16/6/2 (058/2022) and Accounting Research Institute (HICoE) and Ministry of Higher Education, Malaysia for providing the necessary financial assistance for this study.

REFERENCES

[1] S. K. Singh, B. Basuki, and R. Setiawan, "The effect of non-performing loan on profitability: empirical evidence from nepalese commercial banks," *Journal of Asian Finance, Economics and Business*, vol. 8, no. 4, pp. 709–716, 2021, doi: 10.13106/jafeb.2021.vol8.no4.0709.

[2] G. Orlando and R. Pelosi, “Non-performing loans for Italian companies: when time matters. an empirical research on estimating probability to default and loss given default,” *International Journal of Financial Studies*, vol. 8, no. 4, pp. 1–22, Nov. 2020, doi: 10.3390/ijfs8040068.

[3] N. Kuzucu and S. Kuzucu, “What drives non-performing loans? evidence from emerging and advanced economies during pre- and post-global financial crisis,” *Emerging Markets Finance and Trade*, vol. 55, no. 8, pp. 1694–1708, 2019, doi: 10.1080/1540496X.2018.1547877.

[4] K. S. Rajha, “Determinants of non-performing loans: evidence from the jordanian banking sector,” *Journal of Finance and Bank Management*, 2017, doi: 10.15640/jfbm.v5n1a5.

[5] N. Apergis, “Convergence in non-performing loans across EU banks: the role of COVID-19,” *Cogent Economics and Finance*, vol. 10, no. 1, 2022, doi: 10.1080/23322039.2021.2024952.

[6] M. Umar and G. Sun, “Determinants of non-performing loans in Chinese banks,” *Journal of Asia Business Studies*, vol. 12, no. 3, pp. 273–289, 2018, doi: 10.1108/JABS-01-2016-0005.

[7] N. Radivojevic and J. Jovovic, “Examining of determinants of non-performing loans,” *Prague Economic Papers*, vol. 26, no. 3, pp. 300–316, 2017, doi: 10.18267/j.pep.615.

[8] M. Kartikasary, F. Marsintauli, E. Serlawati, and S. Laurens, “Factors affecting the non-performing loans in Indonesia,” *Accounting*, vol. 6, no. 2, pp. 97–106, 2020, doi: 10.5267/j.ac.2019.12.003.

[9] R. K. Asiama and A. Amoah, “Non-performing loans and monetary policy dynamics in Ghana,” *African Journal of Economic and Management Studies*, vol. 10, no. 2, pp. 169–184, 2019, doi: 10.1108/AJEMS-04-2018-0103.

[10] A. Mohamad and H. Jenkins, “Corruption and banks’ non-performing loans: empirical evidence from MENA countries,” *Macroeconomics and Finance in Emerging Market Economies*, vol. 14, no. 3, pp. 308–321, 2021, doi: 10.1080/17520843.2020.1842478.

[11] F. Ofria and M. Mucciardi, “Government failures and non-performing loans in European countries: a spatial approach,” *Journal of Economic Studies*, vol. 49, no. 5, pp. 876–887, 2022, doi: 10.1108/JES-01-2021-0010.

[12] H. Jenkins, E. Alshareef, and A. Mohamad, “The impact of corruption on commercial banks’ credit risk: evidence from a panel quantile regression,” *International Journal of Finance and Economics*, vol. 28, no. 2, pp. 1364–1375, 2023, doi: 10.1002/ijfe.2481.

[13] Y. Desta, “Manifestations and causes of civil service corruption in the of developing countries,” *Journal of Public Administration and Governance*, vol. 9, no. 3, p. 23, 2019, doi: 10.5296/jpag.v9i3.14930.

[14] A. D. Berliani and E. S. Violita, “The role of governance in SDG through public trust in government: study in selected OIC member states,” *IOP Conference Series: Earth and Environmental Science*, vol. 716, no. 1, p. 012100, Mar. 2021, doi: 10.1088/1755-1315/716/1/012100.

[15] T. H. Son, N. T. Liem, and N. V. Khuong, “Corruption, nonperforming loans, and economic growth: international evidence,” *Cogent Business and Management*, vol. 7, no. 1, p. 1735691, Jan. 2020, doi: 10.1080/23311975.2020.1735691.

[16] M. A. Khan, A. Siddique, and Z. Sarwar, “Determinants of non-performing loans in the banking sector in developing state,” *Asian Journal of Accounting Research*, vol. 5, no. 1, pp. 135–145, 2020, doi: 10.1108/AJAR-10-2019-0080.

[17] S. Kepli, Y. Bani, A. Rosland, and N. Laila, “Non-performing loans and macroeconomic variables in Malaysia: recent evidence,” *International Journal of Economics and Management*, vol. 15, no. 1, pp. 19–31, 2021, doi: 10.21315/ijem.2021.15.1.5.

[18] A. Sajari, H. Haron, Y. Ganesan, and A. A. Khalid, “Factors influencing the level of ethics and integrity in Malaysian public sector,” *Journal of Islamic Accounting and Business Research*, vol. 14, no. 1, pp. 141–158, 2023, doi: 10.1108/JIABR-09-2021-0256.

[19] E. M. Hambi and F. Benabbou, “A deep learning based technique for plagiarism detection: a comparative study,” *IAES International Journal of Artificial Intelligence*, vol. 9, no. 1, pp. 81–90, 2020, doi: 10.11591/ijai.v9.i1.pp81-90.

[20] I. Tammouch, A. Elouafi, S. Eddarouich, and R. Touahni, “Centroid competitive learning approach for clustering and mapping the social vulnerability in Morocco,” in *International Journal of Advanced and Applied Sciences (IJAAS)*, vol. 9, no. 9, pp. 70–77, 2022, doi: 10.21833/ijaas.2022.09.009.

[21] L. Barik, “Data mining approach for digital forensics task with deep learning techniques,” *International Journal of Advanced and Applied Sciences (IJAAS)*, vol. 7, no. 5, pp. 56–65, 2020, doi: 10.21833/ijaas.2020.05.008.

[22] N. K. W. Chan, A. S. H. Lee, and Z. Zainol, “A framework for predicting employee health risks using ensemble model,” *International Journal of Advanced and Applied Sciences (IJAAS)*, vol. 8, no. 9, pp. 29–38, 2021, doi: 10.21833/ijaas.2021.09.004.

[23] A. Bellotti, D. Brigo, P. Gambetti, and F. D. Vrins, “Forecasting recovery rates on non-performing loans with machine learning,” *SSRN Electronic Journal*, 2019, doi: 10.2139/ssrn.3434412.

[24] S. I. Serengil, S. Imece, U. G. Tosun, E. B. Buyukbas, and B. Koroglu, “A comparative study of machine learning approaches for non performing loan prediction,” in *Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021*, 2021, pp. 326–331, doi: 10.1109/UBMK52708.2021.9558894.

[25] M. A. Elnaggar, M. A. El-Azeem, and F. A. Maghraby, “Machine learning model for predicting non-performing agricultural loans,” in *Advances in Intelligent Systems and Computing*, 2020, vol. 1153 AISC, pp. 395–404, doi: 10.1007/978-3-030-44289-7_37.

BIOGRAPHIES OF AUTHORS

Associate Professor Ts. Dr. Suraya Masrom is the head of Machine Learning and Interactive Visualization (MaLIV) Research Group at Universiti Teknologi MARA (UiTM) Perak Branch. She received her Ph.D. in Information Technology and Quantitative Science from UiTM in 2015. She started her career in the information technology industry as an Associate Network Engineer at Ramgate Systems Sdn. Bhd. (a subsidiary of DRB-HICOM) in June 1996 after receiving her bachelor’s degree in computer science from Universiti Teknologi Malaysia (UTM) in Mac 1996. She started her career as a lecturer at UTM after receiving her master’s degree in Computer Science from Universiti Putra Malaysia in 2001. She transferred to the Universiti Teknologi MARA (UiTM), Seri Iskandar, Perak, Malaysia, in 2004. She is an active researcher in the meta-heuristics search approach, machine learning, and educational technology. She can be contacted through email at suray078@uitm.edu.my.

Dr. Rahayu Abdul Rahman is an Associate Professor at the Faculty of Accountancy, UiTM. She received her Ph.D. in Accounting from Massey University, Auckland, New Zealand in 2012. Her research interest surrounds areas, like financial reporting quality such as earnings management and accounting conservatism as well as financial leakages including financial reporting frauds and tax aggressiveness. She has published many research papers on machine learning and its application to corporate tax avoidance. She is currently one of the research members of machine learning and interactive visualization research group at UiTM Perak Branch. She can be contacted through email at rahay916@uitm.edu.my.

Nurul Huda Mohd Shukri is a lecturer attached to the Institute of Professional Studies of Universiti Poly-Tech Malaysia, Kuala Lumpur, Malaysia. She holds a Bachelor of Accountancy from UiTM in 2010, Master of Accountancy from UiTM in 2015 and completed her professional accountancy, Association of Chartered Certified Accountants (ACCA), in 2022 from FTMS College. She has been teaching in ACCA and Institute of Chartered Secretaries and Administrators (ICSA) for 8 years and specializing in Audit and Assurance, Financial Reporting, Strategic Business Reporting and also undergraduate programmes. Prior to joining UPTM in 2013, she started her professional career with Deloitte and medium audit firm as audit associate. She is an associate member of the Malaysian Institute of Accountants (MIA) since 2017. His research interest cover topics about financial reporting, governance and CSR disclosure. She can be contacted through email at hudashukri@uptm.edu.my.

Dr Naimah Ahmad Yahya is a Senior Lecturer at the Faculty of Accountancy, Universiti Teknologi MARA (UiTM) Perak Branch. She obtained a Master of Accounting and Information Systems in 2004 from University of South Australia. She received her Ph.D. in Accountancy from UiTM in 2019. Her Ph.D. thesis involves investigation in green intellectual capital and proactive environmental strategy towards manufacturing firms' competitive advantage. Her research interest surrounds areas, like intellectual capital, sustainability accounting, corporate governance and accounting education. She can be contacted through email at naima713@uitm.edu.my.

Mohd Zulfikri Abd Rashid is a lecturer at the Faculty of Accountancy, Universiti Teknologi MARA (UiTM) Perak Branch. He received his master's degree in MBA (Professional Accounting) from Universiti Utara Malaysia in 2011. His areas of research interests include financial accounting and reporting, auditing, corporate governance and taxation. He can be conducted through email at mohdzulfikri@uitm.edu.my.

Dr. Nor Balkish Zakaria is an Associate Professor at the Faculty of Accountancy and senior fellow at Accounting Research Institute, UiTM. She received her Ph.D. in Accounting from Massey University, New Zealand in 2013. Her research interest surrounds areas, like financial reporting quality, earnings management and corporate governance. She can be contacted through email at norbalkish@uitm.edu.my.