Vol.4, No.4, Page 1914- 1923, 2026 https://doi.org/10.59628/jast.v4i4.2121 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

A Comparative Model to Analyze the Impact of Tax Dataset Augmentation on the Accuracy of Machine Learning Models

Abeer Abdullah Shujaaaddeen *, Ammar T.Zahary and Fadl Mutaher Ba-Alwi

Department of Computer Science, Faculty of Computer and Information Technology, Sana'a University, Sana'a, Yemen

*Corresponding author: abeershg@su.edu.ye

Abstract

Many factors influence the efficiency of a classification model in Machine Learning, including the dataset's size and the type of Machine Learning (ML) technique utilized in the classification process. Primarily, the accuracy varies among different machine learning methods. This paper develops a model that analyses and measures the impact of Tax Dataset Augmentation on the Accuracy of Machine Learning Models. The paper compares the performance of different models based on the most common machine learning techniques, namely: DT, RF, SVM, and ANN (MLP). Based on three datasets provided by Yemen's Tax Authority. The first dataset contains 1083 records, while the second dataset is identical to the previous one, but with nearly five times the number of records, following data preprocessing, which resulted in approximately 5000 additional records. The last dataset is the same as the original dataset, but it has been duplicated nearly ten times, resulting in almost 10,000 records. The dataset partitioning technique utilized k-fold validation using the three datasets. Results show that the performance of ML classifiers such as ANN (MLP), DT, RF, and SVM is affected by dataset augmentation in terms of accuracy, recall, precision, and F-score. Results also show that the performance varies among the first three techniques; however, the SVM Classifier yields the lowest results. In general, despite some techniques leading to overfitting, it is found that most machine learning models utilizing tax datasets with five times the duplicates produced better outcomes than those using the original dataset. These findings provide practical guidance for tax authorities in selecting robust machine learning models under limited data availability and highlight risks associated with naïve dataset expansion.

Article Info

Keywords:
ML Techniques, DT, RF, ANN, Performance Measures, Tax Dataset

Article History:
Received: 28-September-2025,
Revised: 01-March-2026,
Accepted: 13-March-2026,
Published: 28 April 2026.

1. INTRODUCTION

Taxes Improving the performance of machine learning models is crucial in the age of data and big data. Model accuracy is a vital component of machine learning applications, directly influencing outcomes and predictions in various fields, including healthcare, trade and tax fraud. The general term "tax fraud" refers to any attempt by organizations or individuals to legally defraud, such as hiding the taxpayer's true status from the tax authorities to lower the tax value. This includes, for instance, filing false tax returns, such as declaring earnings that are below their true value. In other words, tax fraud is the act of lying on a tax return form to lower one's tax obligation. Consequently, one of the top goals of tax authorities is to identify instances of tax fraud [1]. Understanding the relationship between data size and model accuracy can provide a solid foundation for developing more efficient and effective models, which can contribute to enhanced innovation and progress in the field of artificial intelligence.

1.1. Decision Trees

Decision trees (DT) are a supervised machine learning approach for classification, prediction, and feature se-

lection issues. Using the rules learned from the given dataset, it seeks to forecast the target class [2].

1.2. Random Forest

Random Forest (RF) is an ensemble algorithm. This algorithm utilizes a combination of numerous Decision Trees (DT), similar to a forest composed of many trees (Figure 7). A DT has nodes and branches, and the nodes decide whether to remain on a specific branch. The DT determines which class to assign to an object by making successive selections that consider all the features. The RF considers the choices made by a given K number of trees [3].

1.3. Support Vector Machine

One popular machine learning classifier is the Support Vector Machine (SVM). It is used in real-world domains for linear and nonlinear problems. In SVM, instances of classes are separated by a hyperplane. SVM is ideally suited for non-linear classification issues because of its kernel function, which transforms a low-dimensional space into a high-dimensional space. In summary, SVM can be applied to the classification of instances in challenging situations [3].

1.4. Multi-Layer Perceptron

A multi-layer perceptron (MLP) is a member of the feed-forward artificial neural network (ANN) class. An artificial neural network (ANN) simulates human brain function. The way the brain takes in information, processes it, and generates output is the primary source of inspiration for ANN. A perceptron is a fundamental component of an ANN. Every perceptron uses an activation function to generate an output and has an attached weight value. An ANN operates by using training data to learn representations and then connecting them to the desired output variable. Data compression, character recognition, computer vision, pattern recognition, and robotics are some of the real-world applications of artificial neural networks [4].

This study aimed to develop a new model that focuses on analyzing the impact of doubling the tax dataset on the accuracy of machine-learning models. Accordingly, this study aims to (1) analyze the impact of dataset multiplication on ML performance, (2) compare four classifiers under varying data volumes, and (3) identify overfitting risks in tax datasets.

This paper is organized into five sections. The remainder of this paper is organized as follows.

Section 2 presents related work. Section 3 presents the Proposed Methodology. Section 4 presents the experiments. Section 5 presents the General Discussion. Section 6 presents the Conclusion, Future Work, and References.

2. RELATED WORK

Many studies have addressed the use of artificial intelligence and machine learning models to detect financial and tax evasion.

Using personal income tax returns (IRPF, in Spanish) filed in Spain, the study in [5] employed Multi-Layer Perceptron (MLP) neural network models and strong machine learning predictive techniques to assist in detecting tax fraud.

The authors of [6] had two distinct scientific objectives: They began by trying to find out how SMEs perceived the current situation and the steps that needed to be taken to cut down on the related red tape. Second, they attempted to establish a link between tax burdens and entrepreneurship using hierarchical cluster analysis and descriptive statistics.

The authors of [7] used a performance model and an expert system to suggest an abstract solution for tax evasion.

The possible advantages of tax authorities utilizing their operational data were examined by [8]. The purpose of this study was to determine which machine learning techniques are effective in detecting a certain type of fraud. To use the data to uncover bank fraud, researchers used data mining techniques [9].

The authors of [10] proposed a prediction engine that was validated on a dataset of tax defaults and non-defaults at limited liability companies in Finland.

The authors of [11] focused on tax fraud detection. The authors argued that the few labeled data points (known fraud/legal cases) were not indicative of the population as a whole because of sample selection bias. The authors used methods for unsupervised anomaly detection.

The authors of [12] examined the research corpus on new technology audits and taxes. They also provide a future research plan for this topic. In this study, the researchers used artificial intelligence, big data, and blockchain technology.

This effort was motivated primarily by the use of machine learning to help make decisions regarding service taxes in fiscal audit plans for the municipality of São Paulo [13].

The authors of [14] attempted to determine the function of AI in the Indian tax system by considering factors such as the tax system's complexity, tax education, legal fines, and relationships with tax officials.

As a decision support system, the authors of [15] proposed integrating trigger data from taxpayers with predictive analytics using machine learning. This study makes it easier to develop predictive analytics algorithms that can precisely identify potential taxpayers who are most likely to pay their fair share.

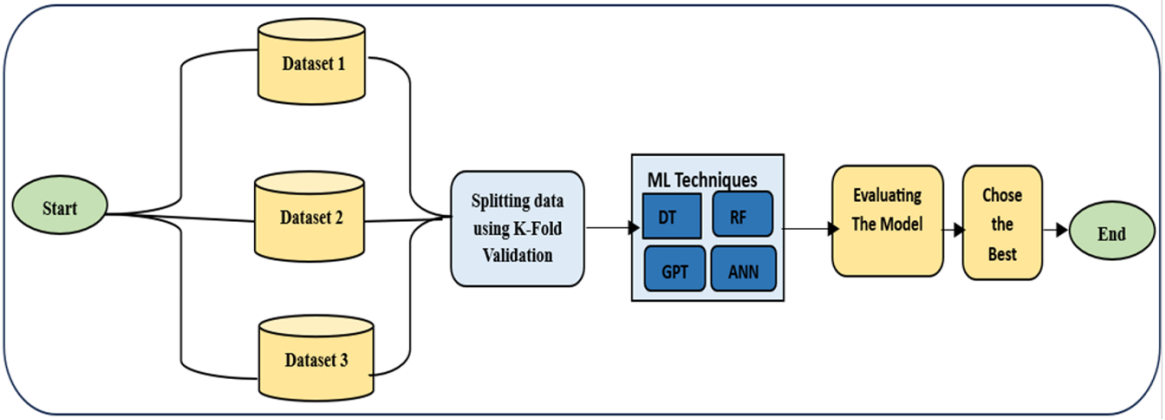

Figure 1. The Proposed Model

The authors of [16] aimed to better identify tax evasion by examining the impact of wealth in Lithuania using data mining technology.

Finally, the primary objective of the study conducted in [17] was to develop a framework for detecting tax fraud using a supervised, unsupervised, behavioral, and predictive modules.

Unlike our previous studies [18–20], which focused on data segmentation strategies and classifier comparisons, this study uniquely focuses on the sensitivity of classifiers to the dataset size scaling. It compares classifiers before and after data scaling and explicitly analyzes the risks of over-allocation in resource-constrained environments.

Despite extensive studies on tax fraud detection, limited attention has been paid to the effect of artificial dataset expansion on classifier reliability, particularly in developing countries. This gap motivated the present study.

3. PROPOSED METHODOLOGY

Researchers employed a set of Machine Learning (ML) algorithms to investigate the impact of data multiplication on the effectiveness of machine learning models. The researchers used three datasets: a tax dataset consisting of 1,083 records, which they multiplied five times at one time and ten times at another time to create an expanded dataset. The datasets were divided using cross-validation, and K was set to ten. The methodology of the proposed study is shown in Figure 1. The proposed study design was based on a supervised machine-learning approach.

3.1. Dataset

The researchers collected the tax dataset after preprocessing the data. The dataset was updated to include commercial and industrial profit taxes. A total of 1083 records were collected. The data are described along with the definitions of the fields and their influence on other factors from a tax accounting standpoint, as shown in Table 1.

Target Variable Definition:

The target classification variable represents tax compliance status. Labels were derived from discrepancies between the declared tax and computed obligations using the “Due Tax” and “Deserved Amount” fields, following Yemeni Tax Authority auditing rules. Records representing taxpayer characteristics were classified into multiple categories or levels based on the degree of tax evasion (total evasion, partial evasion, and compliance), allowing for multi-classification.

Separate papers published in 2023 and 2024 [1][18][19] by the same authors explain each preprocessing step. Subsequently, the researchers duplicated the dataset twice. First, the dataset was duplicated five times from the original dataset, and the second time, the dataset was duplicated ten times from the same original dataset to create new datasets in addition to the original dataset.

Dataset Evaluation Strategy:

All experiments were conducted using stratified 10-fold cross-validation exclusively to ensure robustness with a small sample size. Earlier references to a 70/30 train–test split were removed to avoid ambiguity. Cross-validation was used to reduce variance and provide stable performance estimates.

Splitting the Dataset:

The dataset was split into two sets: 30% for testing and 70% for training. Owing to data imbalances, the researchers trained the data using k-fold cross-validation. The algorithms used in the experiments included Multilayer Neural Network (MLP ANN), support vector machine (SVM), Decision Tree (DT), and Random Forest (RF)..

Ethical Considerations and Data Governance:

The dataset was officially provided by the Yemeni Tax Authority through an institutional collaboration. All identifiers were anonymized prior to analysis. The data were used solely for academic research purposes and

Table 1. The variables of Taxes

No	Type	The description	Name of variable
1	Integer/number	Tax Number	TIN
2	Var Char	Trade Name	TN
3	Var Char	legal entity	LE
4	Integer/number	Tax period	TP
5	Var Char	Tax type	T_type
6	Integer/number	Business Number	BN
7	Integer/number	Tax	Tax
8	Integer/number	Payable under the account	Punder
9	Integer/number	Due tax	Dtax
10	Integer/number	Fines	Fine
11	Integer/number	Deserved amount	Damount
12	Integer/number	Tax rate to turnover	Tax_L div BN_L
13	Integer/number	Tax rate to turnover	Tax_C div BN_C
14	Integer/number	Tax rate on turnover for the previous year	Ratio_C

complied with the local data protection regulations.

3.2. ML Techniques

The types of Machine Learning (ML) techniques used in this study were Decision Trees (DT), Random Forest (RF), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). Table 2 presents the hyperparameter settings.

Table 2. Hyperparameter Settings

RF	No-trees = 30, max_depth = 10, criterion = Gain Ratio
DT	max_depth = 10, criterion = Gain Ratio
SVM	C-SVC, Sigmoid kernel
ANN	4 layers, 10 neurons, ReLU (hidden), SoftMax (output), epochs = 10

3.3. Evaluation of The Model

The model was evaluated using a Confusion Matrix. The confusion matrix, which measures the efficacy of a classification model, is an N × N matrix, where N is the number of target classes in the model. Visual representation of the confusion matrix. The correctness of the model was determined by examining the diagonal values and counting the correct categories. The confusion matrix is a square matrix with the predicted values for the model shown in the row and the actual values in the column.

TP occurs when a positive value is predicted by the model and a positive outcome is obtained.

FP (False Positive): Despite being accurate, the forecast was incorrect. (A Type 1 error).

FN (False Negative): The prediction and outcome are both incorrect. This is referred to as Type 2 error.

TN (True Negative): both the actual value and the model's forecast are negative [20].

Table 3 presents the performance metrics of our model.

4. RESULTS OF EXPERIMENTS

The researchers executed many experiments with four ML techniques (RF, DT, SVM, and ANN) using three datasets: one with 1083 records, the second with 5000 records, and the last with 10.000 records, to compare the results and then choose the best techniques. They used k-fold validation and set K=10 with shuffled samples used in the sampling process.

4.1. RF Model Results

When the researchers used a dataset of 1083 records with RF and the hyperparameter number of trees = 30, the criterion was the gain ratio, and max depth was 10, the results were as follows: they achieved 96.56% accuracy,66.67% recall,56.54% in precision, and 61.18% in f-score. When we used a dataset of 5.000 records with the same technique, it achieved 97.40% accuracy, 66.67% recall, 65.67% precision, and 66.16% F-score. When we used a dataset of 10.000 records with the same technique, it achieved 97.30% accuracy, 66.67% recall, 65.66% precision, and 66.16 % F-score, as shown in Table 4 and Fig. 2.

4.2. DT Model Results

When the researchers used a dataset of 1083 records with DT as the hyperparameter, the criterion was the gain ratio, and the maximum depth was 10. The results were as follows: it achieved 96.42% accuracy, 63.33% recall, 62.12% in precision, and 62.71% in F-score. When we used a dataset of 5.000 records with the same technique, it achieved 97.07% accuracy, 66.67% recall, 65.66% precision, and 60.22% F-score. When we used a dataset of 10.000 records with the same technique, it achieved 97.09% accuracy, 66.67% recall, 65.65% precision, and 66.15% F-score, as shown in Table 5 and Fig. 3.

Table 3. Performance measures

N	Measure	Description	Equation
1	Accuracy	Accuracy is the number of correct predictions the model makes for the full test dataset. Accuracy is a useful metric to evaluate the model's performance in balanced datasets. Accuracy is a bad statistic with unbalanced datasets [1].	$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ (1)
2	Precision	Precision indicates the proportion of accurately predicted events that occurred and were advantageous. This will show whether the model is trustworthy or not. Precision is a helpful indicator when a False Positive is more problematic than a False Negative [15][21][22].	$Precision = \frac{TP}{TP+FP}$ (2)
3	Recall	The number of real positive cases that our model was able to predict is indicated by recall. When there are false-positive or false-negative outcomes, the recall measure is useful [1] [19].	$Recall = \frac{TP}{TP+FN}$ (3)
4	F-Score	The mean of precision and recall is known as the F-measure (F-score). It is a performance metric that integrates precision and recall into one. The F1-score is equally influenced by precision and recall [13][21].	$F\text{-}score = \frac{2 \times (Precision \times Recall)}{Precision + Recall}$ (4)

Table 4. The result of the performance measure for RF

Measure	Training		
RF	Dataset=1083	Dataset=5000	Dataset=10000
Accuracy	96.56%	97.40%	97.30%
Recall	66.67%	66.67%	66.67%
Precision	56.54%	65.67%	65.66%
F-score	61.18%	66.16%	66.16%

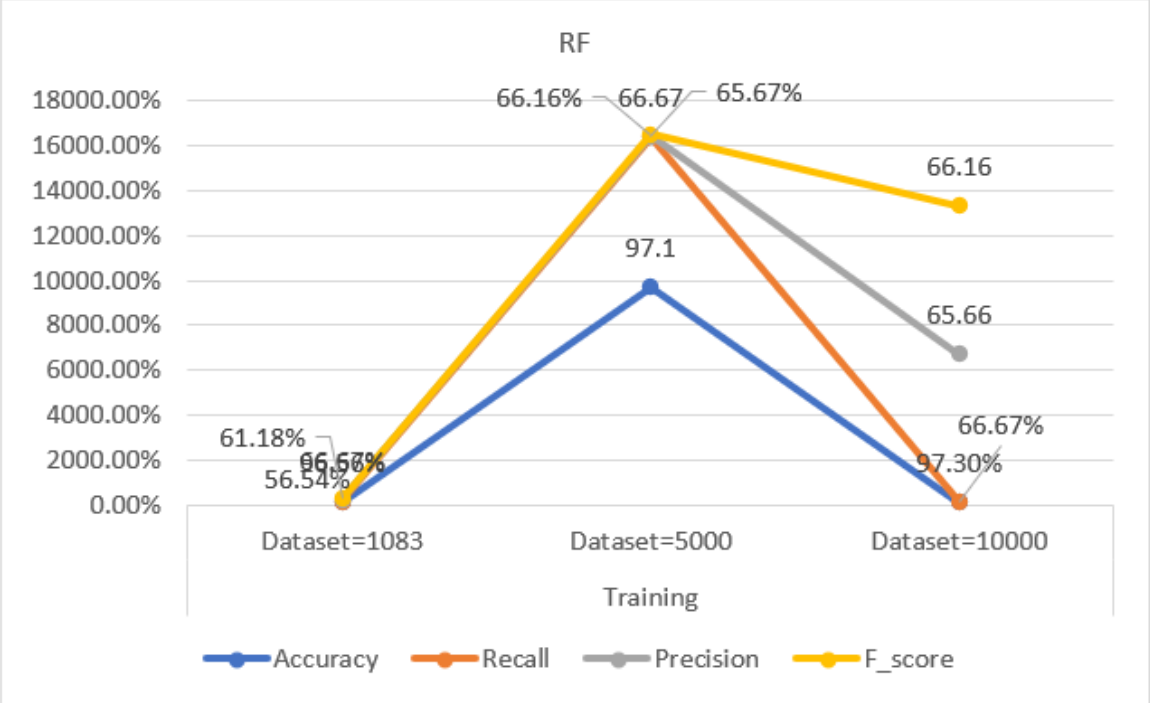

Figure 2. Confusion Matrix Chart for RF

Table 5. The result of the performance measure for DT

Measure	Training		
DT	Dataset=1083	Dataset=5000	Dataset=10000
Accuracy	96.42%	97.07%	97.09%
Recall	63.33%	66.67%	66.67%
Precision	62.12%	65.66%	65.65%
F-score	62.71%	60.22%	66.15%

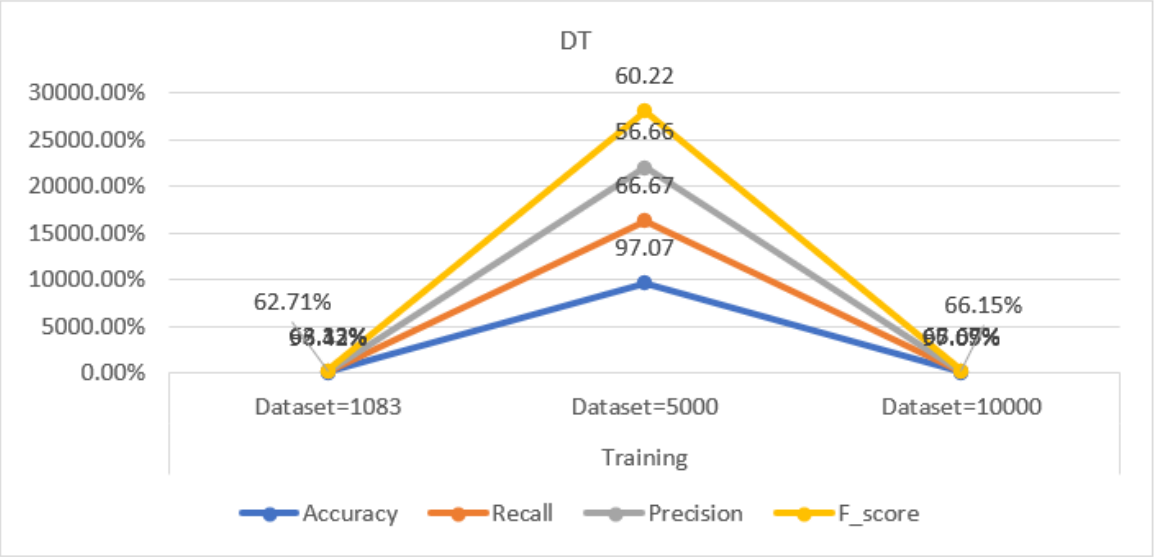

Figure 3. Confusion Matrix for DT

4.3. SVM Model Results

When the researchers used a dataset of 1083 records with SVM with the hyperparameters SVM type = C-SVC and Kernel type = Sigmoid, the results were as follows: SVM achieved 94.12% accuracy, 33.29% recall, 32.27% precision, and 32.7% F-score. When we used a dataset of 5.000 records with the same technique, it achieved 96.54% accuracy, 66.54% recall, 65.60% precision, and 66.05% F-score. When we used a dataset of 10.000 records with the same technique, it achieved 100% in all measures, as shown in Table 6 and Fig. 4.

4.4. ANN (MLP) Model Results

The researchers used a dataset of 1083 records with ANN (MLP) with the hyperparameter many layers = 4 (one input layer, two hidden layers, and one output layer), number of neurons =10 on hidden layers, activation function in hidden layers = ReLU, activation function in output layer = SoftMax, and number of epochs = 10, as shown in Fig. 5.

The results were as follows: the technique ANN (MLP) achieved 98.96% accuracy, 89.88% recall, 86.15% precision, and 87.97% f-score. When we used a dataset of 5.000 records with the same technique, it achieved 99.96% accuracy,99.99% recall,99.64% Precision, and 99.81% F-score. When we used a dataset of 10.000 records with the same technique, it achieved 100 % in all measures, as shown in Table 7 and Fig. 6.

5. GENERAL DISCUSSION

In this study, we used four evaluation metrics to evaluate the performance of four machine learning models, namely RF, DT, SVM, and ANN (MLP), using three data sets. The four metrics were accuracy, precision, recall, and F-score. Using a dataset of 1083 records, the results achieved 98.96% accuracy, 89.88% recall, 86.15% precision, and 87.97% F-score by the ANN (MLP). The RF technique achieved 96.56% accuracy, 66.67% recall, 56.54% precision, and 61.18% F-score. The DT classifier has achieved 62.71% F-score, 62.12% precision, 63.33% recall, and 96.42% accuracy. In contrast, the SVM Classifier provided the worst results; it achieved 94.12% accuracy, 33.29% recall, 32.27% precision, and 32.7% F-score.

When the researchers used a dataset of 5000 records with ANN, it achieved 99.96% accuracy, 99.99% recall, 99.64% precision, and 99.81% F-score; then, 97.10% accuracy, 66.67% recall, 65.67% precision, 66.16% f-score with RF; after that, 97.07% accuracy, 66.67% recall, 56.66% precision, 60.22% f-score with DT. On the

Table 6. The result of the performance measure for SVM

Measure	Training		
SVM	Dataset=1083	Dataset=5000	Dataset=10000
Accuracy	94.12%	96.54%	100%
Recall	33.29%	66.54%	100%
Precision	32.27%	65.60%	100%
F-Score	32.7%	66.05%	100%

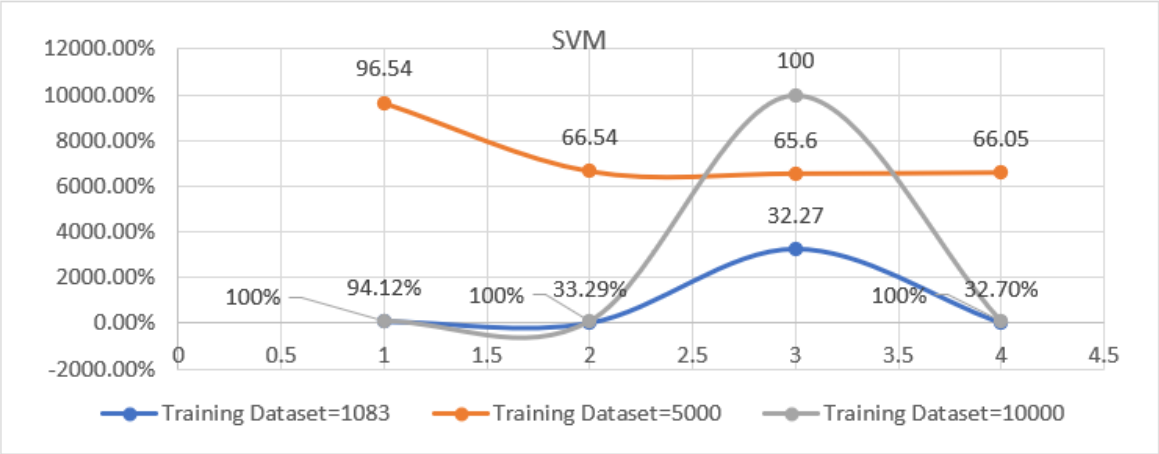

Figure 4. Confusion Matrix Chart for SVM

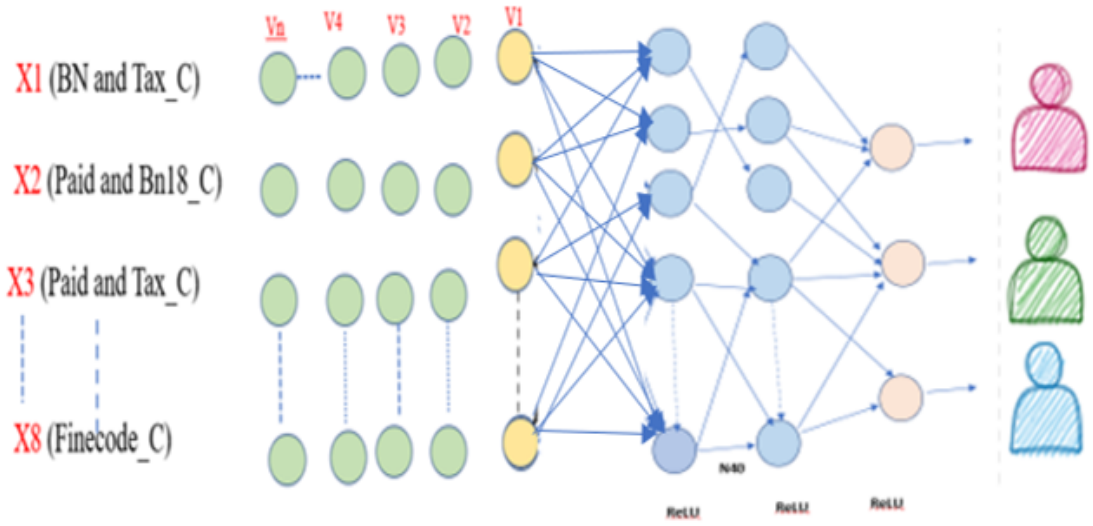

Figure 5. ANN Network for the tax dataset

Table 7. The result of the performance measure for ANN

Measure	Training		
ANN(MLP)	Dataset=1083	Dataset=5000	Dataset=10000
Accuracy	98.96%	99.96%	100%
Recall	89.88%	99.99%	100%
Precision	86.15%	99.64%	100%
F-score	87.97%	99.81%	100%

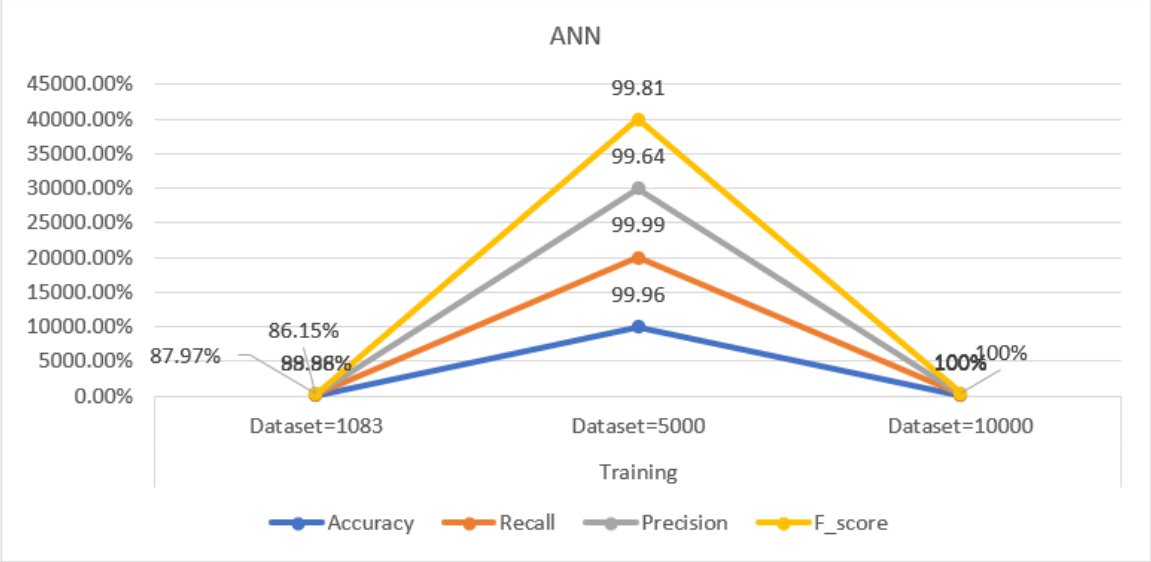

Figure 6. Confusion Matrix Chart for ANN

other hand, when we used SVM, it achieved the lowest scores: 96.54% accuracy, 66.54% recall, 65.60% precision, and 66.05% F-score. When we used a dataset of 10000 records, they achieved the highest scores in ANN and SVM, with 100% in all the measures, but these results led to overfitting. RF provided 97.30% accuracy, 66.67% recall, 65.66% precision, and 66.16% F-score. When we used DT, it achieved 97.09% accuracy, 66.67% recall, 65.65% precision, and 66.15% F-score, as shown in Table 8 and Fig. 7.

Limitations of Dataset Duplication:

Dataset expansion in this study relied on direct record duplication, which did not introduce new variability. This approach was intentionally used as a stress test to examine the model sensitivity to the sample volume rather than as a genuine augmentation. Consequently, the perfect scores observed in ANN and SVM on the 10,000-record dataset indicate overfitting caused by the redundancy. Future studies will adopt synthetic augmentation methods, such as SMOTE and noise-based generation, combined with grouped cross-validation.

ANN achieved superior performance owing to its ability to model the nonlinear relationships inherent in financial variables, whereas SVM showed instability under limited sample sizes. This behavior aligns with prior findings on imbalanced fiscal datasets and highlights the importance of model selection in realistic tax environments.

Statistical significance testing was not conducted because of dataset constraints; future work will incorporate bootstrapping and nested validation for a stronger inference.

6. CONCLUSION AND FUTURE WORK

In this study, the performance of a series of machine learning models, DT, RF, SVM, and ANN (MLP), was compared using three datasets: the original dataset provided by the Yemeni Tax Authority and two duplicated datasets. Using the original dataset of 1083 records, the researchers achieved 98.96% accuracy,89.88% recall,86.15% precision, and 87.97% F-score with the ANN. Then, The RF classifier achieved 61.18% F-score, 56.54% precision, 66.67% recall, and 96.56% accuracy. DT achieved 96.42% accuracy,63.33% recall, 62.12% precision, and 62.71% F-score. In contrast, the SVM Classifier provided the lowest results, with 94.12% accuracy, 33.29% recall, 32.27% precision, and 32.7% F-score. Conversely, the ANN achieved 98.96% accuracy, 89.88% recall, 86.15% precision, and 87.97% F-score.

When the researchers used the dataset of 5000 records with ANN, they achieved

99.96% accuracy,99.99% recall,99.64%

precision, and 99.81% F-score; then 97.40% accuracy, 66.67% recall, 65.67% precision, 66.16% f-score with RF; after that, 97.07% accuracy, 66.67% recall, 56.66% precision, 60.22% f-score with DT. On the other hand, when we used SVM, it achieved the lowest scores: 96.54% accuracy, 66.54% recall, 65.60% precision, and 66.05% F-score. When they used a dataset of 10.000 records, they achieved the highest scores in ANN and SVM with 100% in all the measures, but these results led to overfitting. RF achieved 97.30% accuracy, 66.67% recall, 65.66% precision, and 66.16% F-score. When we used DT, it achieved 97.09% accuracy, 66.67% recall, 65.65% precision, and 66.15% F-score. As future work, we propose applying our model to larger datasets in other financial and tax fields, such as sales tax and different types of taxes.

Table 8. The result of the performance measure

Classifier	Dataset = 1083				Dataset = 5000				Dataset = 10,000			
	Acc.	Rec.	Prec.	F1	Acc.	Rec.	Prec.	F1	Acc.	Rec.	Prec.	F1
DT	96.42	63.33	62.12	62.71	97.09	66.67	65.66	60.22	97.09	66.67	65.65	66.15
RF	96.56	66.67	65.54	61.18	97.40	66.67	65.67	66.16	97.30	66.67	65.66	66.16
SVM	94.12	33.2	32.27	32.7	96.54	66.54	65.60	66.05	100	100	100	100
ANN	98.96	89.88	86.15	87.97	99.96	99.99	99.64	99.81	100	100	100	100

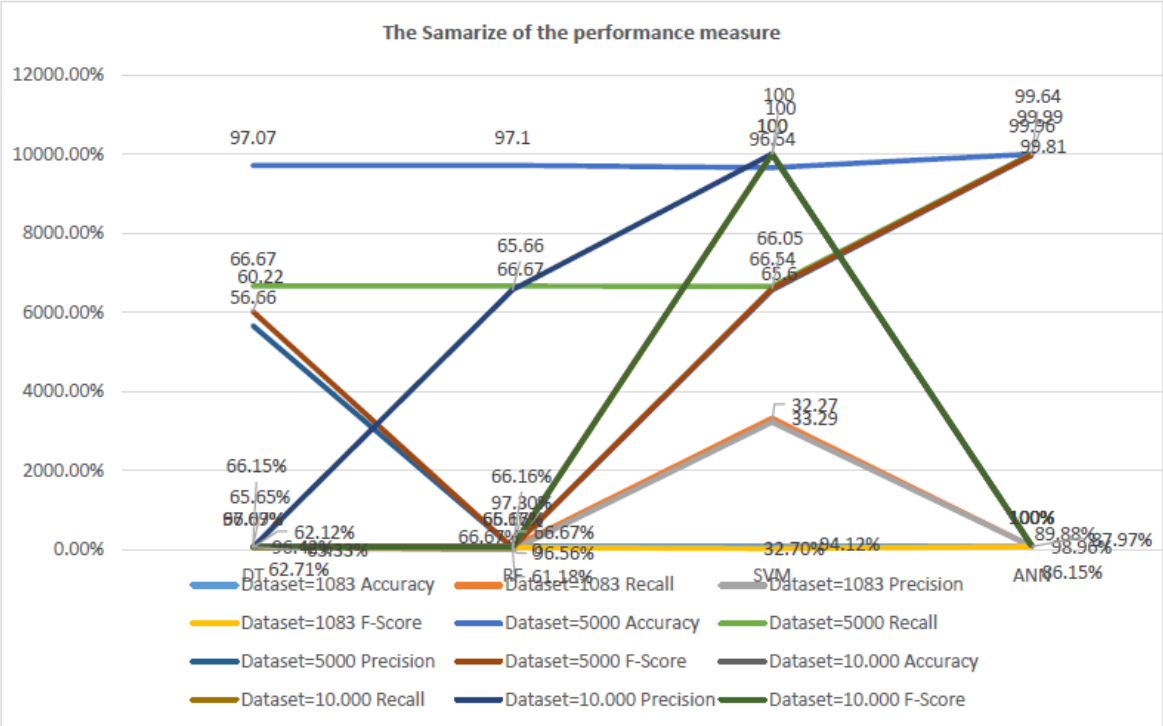

Figure 7. Confusion Matrix Chart for all ML techniques

Practically, these results suggest prioritizing ANN models only when sufficient genuine data are available, whereas Random Forest offers more stable behavior under limited samples. Tax authorities are advised to focus on real data acquisition rather than artificial duplication.

REFERENCES

[1] A. Shujaaddeen, F. M. Ba-Alwi, and G. Al-Gaphari, "A new machine learning model for detecting levels of tax evasion based on hybrid neural network," *Int. J. Intell. Syst. Appl. Eng.*, vol. 12, no. 11s, pp. 450–468, Jan. 2024.

[2] K. A. Sadi, "Prediction model of type 2 diabetes mellitus for oman prediabetes patients using artificial neural network and six machine learning classifiers," *Appl. Sci.*, 2023.

[3] M. T. Abraham, N. Satyam, P. Jain, B. Pradhan, and A. Alamri, "Effect of spatial resolution and data splitting on landslide susceptibility mapping using different machine learning algorithms," *Geomatics, Nat. Hazards Risk*, vol. 12, no. 1, pp. 3381–3408, 2021. DOI: 10.1080/19475705.2021.2011791.

[4] A. A. Shujaaddeen, F. M. Ba-Alwi, A. T. Zahary, A. S. Alhegami, A. Alsabry, and A. M. Al-Badani, "A binary and multi classification model on tax evasion: A comparative study," in *2024 1st International Conference on Emerging Technologies for Dependable Internet of Things (ICETI)*, Sana'a, Yemen, 2024, pp. 1–9. DOI: 10.1109/ICETI63946.2024.10777224.

[5] D. Rodr, "Tax fraud detection through neural networks: An application using a sample of personal income taxpayers," *Future Internet*, vol. 11, no. 4, p. 86, 2019. DOI: 10.3390/fi11040086.

[6] D. Ravšelj, P. Kovač, and A. Aristovnik, "Tax-related burden on smes in the european union: The case of slovenia," *Soc. Sci.*, vol. 10, no. 2, 2019.

[7] D. Bogdanović and E. Babović, "Expert systems as a means in detecting tax evasion," 2020.

[8] A. Z. Adamov, "Machine learning and advanced analytics in tax fraud detection," in *2019 IEEE 13th International Conference on Application of Information and Communication Technologies (AICT)*, IEEE, 2019, pp. 1–5.

[9] V. Kavitha, S. Krithika, M. Tejaswini, and N. Nihitha, "An income tax fraud detection using ai," *J. Crit. Rev.*, vol. 7, no. 16, pp. 119–124, 2020.

[10] M. Z. Abedin, G. Chi, M. M. Uddin, M. S. Satu, M. I. Khan, and P. Hajek, "Tax default prediction using feature transformation-based machine learning," *IEEE access*, vol. 9, pp. 19 864–19 881, 2020.

[11] V. Jellis, M. David, P. Bruno, J. Vanhoeyveld, D. Martens, and B. Peeters, "Value-added tax fraud detection with scalable anomaly detection techniques," *Arch. Author Version (Peer-reviewed)*, vol. 86, 2020.

[12] O. F. Atayah, "Audit and tax in the context of emerging technologies: A retrospective analysis, current trends, and future opportunities," *Int. J. Digit. Account. Res. (assumed)*, vol. 21, pp. 95–128, 2021. DOI: 10.4192/1577-8517-v21.

[13] A. Ippolito and A. C. G. Lozano, "Tax crime prediction with machine learning: A case study in the municipality of são paulo," in *Proceedings of the 22nd International Conference on Enterprise Information Systems (ICEIS)*, vol. 1, 2020, pp. 452–459. DOI: 10.5220/0009564704520459.

[14] A. Rathi, S. Sharma, G. Lodha, and M. Srivastava, "A study on application of artificial intelligence and machine learning in indian taxation system," *J. Physics: Conf. Ser. / PAE (uncertain)*, 2021. DOI: 10.17762/pae.v58i2.2265.

[15] R. D. Febriminanto and M. Wasesa, "Machine learning for predicting tax revenue potential," *Indonesian Treas. Rev. J. Perbendaharaan, Keuangan Negara dan Kebijakan Publik*, 2022, Available online: www.pajak.com.

[16] T. Ruzgas, L. Kižauskienė, M. Lukauskas, E. Sinkevičius, M. Frolovaitė, and J. Arnastauskaitė, "Tax fraud reduction using analytics in an east european country," *Axioms*, vol. 12, no. 3, p. 288, Mar. 2023. DOI: 10.3390/axioms12030288.

[17] N. Alsadhan, "A multi-module machine learning approach to detect tax fraud," *Comput. Syst. Sci. Eng.*, vol. 46, no. 1, pp. 241–253, 2023. DOI: 10.32604/csse.2023.033375.

[18] A. A. Shujaaddeen, F. M. Ba-Alwi, A. T. Zahary, and A. S. Alhegami, "A model for measuring the effect of splitting data method on the efficiency of machine learning models: A comparative study," in *2024 4th International Conference on Emerging Smart Technologies and Applications (eSmarTA)*, Sana'a, Yemen, 2024, pp. 1–13. DOI: 10.1109/eSmarTA62850.2024.10639022.

[19] A. A. S. Shujaaddeen and F. M. M. Ba-Alwi, "A comparative study of the performance of machine learning models on a tax dataset of yemen to detect levels of tax evasion," *Sana'a Univ. J. Appl. Sci. Technol.*, vol. 1, no. 4, pp. 304–312, 2023. DOI: 10.59628/just.v1i4.528.

[20] A. A. Shujaaddeen, F. M. Ba-Alwi, A. T. Zahary, G. Al-Gaphari, A. M. Al-Badani, and A. Alsabry, "Enhancing a random forest model based on single rule reduction for tax evasion depends on the values of k in k-fold validation technique," in *2024 1st International Conference on Emerging Technologies for Dependable Internet of Things (ICETI)*, Sana'a, Yemen, 2024, pp. 1–9. DOI: 10.1109/ICETI63946.2024.10777271.

[21] J. Brownlee, *How to calculate precision, recall, and f-measure for imbalanced classification*, Accessed: 2026-04-10, Jan. 2020. [Online]. Available: https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalance.

[22] A. Kulkarni, "Confusion matrix," *ScienceDirect*, pp. 1–22, 2022.