An evaluation of nature-inspired optimization algorithms and machine learning classifiers for electricity fraud prediction

Ami Shamril Kamaruddin[1], Mohd Fikri Hadrawi[2], Yap Bee Wah[3], Sharifah Aliman[1]
[1]School of Computing Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia
[2]School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, Shah Alam, Malaysia
[3]UNITAR Graduate School, UNITAR International University, Petaling Jaya, Malaysia

Article Info

Article history:

Received Jan 14, 2023
Revised Jul 12, 2023
Accepted Jul 15, 2023

Keywords:

Artificial neural network
Electricity fraud
Extreme gradient-boosted tree
Grey wolf optimizer
Imbalance class
Particle swarm optimization
Support vector machine

ABSTRACT

This study evaluated the nature-inspired optimization algorithms to improve classification involving imbalanced class problems. The particle swarm optimization (PSO) and grey wolf optimizer (GWO) were used to adaptively balance the distribution and then four supervised machine learning classifiers artificial neural network (ANN), support vector machine (SVM), extreme gradient-boosted tree (XGBoost), and random forest (RF) were applied to maximize the classification performance for electricity fraud prediction. The imbalance data was balanced using random undersampling (RUS) and two nature-inspired algorithm techniques (PSO and GWO). Results showed that for the data balanced using random undersampling, ANN (Sentest = 50.31%), and XGBoost (Sentest = 66.32%) has better sensitivity than SVM (Sentest = 23.61%), while RF exhibits overfitting (Sentrain = 100%, Sentest = 71.25%). The classification performance of RF model hybrid with PSO improved tremendously (AccTest = 96.98%, Sentest = 94.87%, Spectest = 99.16%, Pretest = 99.14%, F1 Score = 96.96%, and area under the curve (AUC) = 0.989). This was closely followed by hybrid of XGBoost with PSO. Moreover, RF and XGBoost hybrid with GWO also showed an improvement and promising results. This study has showed that nature-inspired optimization algorithms (PSO and GWO) are effective methods in addressing imbalanced dataset.

Corresponding Author:

Mohd Fikri Hadrawi
School of Mathematical Sciences, College of Computing, Informatics and Mathematics
Universiti Teknologi MARA
40450 Shah Alam, Selangor, Malaysia
Email: fikri1814@uitm.edu.my

1. INTRODUCTION

The electricity losses are commonly categorized into two categories: technical losses and non-technical losses (NTL). The NTL are primarily due to fraud caused by meter manipulating or bypassing, fake meter readings, broken meters, or un-metered supply. Glauner [1] reported NTL was range up to 40% from the total electricity distributed. Electricity fraud refers to intentional and illegal usage of electricity by various means. The simplest and most common way of electricity fraud is by connecting directly to energy sources bypassing the metering process tampering the meter reading, tempering the firmware or storage of the smart meters, interrupting communications, interfering measurements, or by modifying the data by gaining unauthorized access to the smart meter [2]. Electricity fraud is a serious issue and the main challenge faced by the electricity provider and has profound effects such as financial losses, the ability to invest in future development, provide stable services to their customers, and indirectly will increase the electricity prices. Tenaga Nasional Berhad (TNB) lost RM106 million from electricity fraud activities in 2019 with 457 cases in the peninsular where Selangor recorded the highest number of electricity fraud [3]. Most countries have been substituting their old electrical system into smart grid in line with advance technology of information system and modern communication. Smart grid is two-way transmission and communication for energy and data respectively. Advanced metering infrastructure (AMI) is one key technology being used currently in smart grid where the old mechanical electricity metering systems have been modernized by smart meters. With these smart meters, electricity consumption data will be transmitted to a control centre at certain time intervals.

In most studies, fraud prediction was faced with technical issue which is imbalanced dataset [4] where the minority class is less than 5%. It is considered as a rare event, and it is tough to get an essential and decent predictive model because of limited information to train from the rare event [5]. The imbalanced data problem is mostly addressed using data level approach such as oversampling the minority class or under sampling the majority class which may cause issues such as overfitting or underfitting problem and most approaches focus on generating a synthetic fraud data such as random oversampling and the synthetic minority over sampling technique (SMOTE). In recent years, nature-inspired algorithms have gained attention in solving imbalanced data problem. Nature-inspired optimization algorithms have been shown to perform well in solving real world optimization problems in various industries, from engineering, medical, financial, industrial, and educational research. ANT colony, artificial bee colony and bat algorithm and particle swarm optimization (PSO) are some conventional natured-inspired algorithms. In recent years, new algorithm such as grey wolf optimization (GWO) was introduced by Mirjalili *et al.* [6]. Thus, this study investigates the algorithmic approach based on nature-inspired optimization algorithms to minimize the effect of imbalanced dataset problem and improve the machine learning classifiers performance in electricity fraud prediction. The goal of this study is to optimize electricity fraud prediction by utilizing nature-inspired optimization algorithms and machine learning classifier which can be deployed by electricity providers.

2. METHOD

2.1. Data description

Electricity fraud dataset consists of daily electricity consumption data of 33841 customers in 1033 days starting from Jan 2014 till Oct 2016. Out of 33841 records, 3615 are flagged as electricity fraud cases. The description of the variables is given in Table 1. The target variable is Electricity Fraud (fraud = 1 or non-fraud = 0) and the predictor variable is customer electricity consumption (kWh).

Table 1. Description of variables

Variable Name	Role	Variable Type	Description
Electricity Fraud (EF)	Target (Y)	Binary	1: fraud 0: non-fraud
Electricity Usage (EU)	Input (X)	Continuous	Daily electricity consumption in kWh

2.2. Data pre-processing

There are almost 26% (9,013,278) missing values in the dataset. According to [7], there is no established cutoff from the literature regarding an acceptable percentage of missing data for valid statistical inferences. Thus, customers whose have more than 50% missing values been omitted. The missing value of the dataset is now reduced to 9.79%. Data imputation was performed by replacing the remaining missing values with the consumption values from the day before or day after. There are ten customers with highest electricity consumption considered as outliers and were omitted.

The cleaned electricity dataset consists of 23849 customer records is imbalanced with only 2402 (10.07%) fraud samples and 21447 (89.93%) non-fraud samples. Li *et al.* [8] reported PSO has performed effectively for imbalanced data in health and medical datasets regardless of dataset sizes. Also, on the hybrid level approach by combining these approaches, Haya [9] applied SMOTE, random undersampling and optimizing the C, γ and kernel type of support vector machine (SVM), and reported 96% accuracy for class imbalance in a direct marketing dataset. For comparison purpose, this imbalanced data was balanced using Random undersampling (RUS) and two nature-inspired algorithm techniques (PSO and GWO) to undersample the majority class. Then the performance of artificial neural network (ANN). SVM, random forest (RF) and extreme gradient-boosted tree (XGBoost), were evaluated using the three balanced datasets plus the original imbalanced dataset.

Four machine learning models which are ANN, SVM, XGBoost, and RF were developed using the cleaned imbalanced dataset ($Data_{ORI}$) and the three balanced datasets ($Data_{RUS}$, $Data_{PSO}$ and $Data_{GWO}$). Data was partitioned by splitting the data randomly into training (80%) and testing (20%) samples. Then, machine classifiers (ANN, SVM, RF, XGBoost) were developed using the train samples (70%) and evaluated based on the testing samples (30%) for each of the four datasets.

2.3. Machine learning models

Machine learning is a technique that use statistical models which provides the ability to learn patterns from available data with intention to make predictions [1]. Machine learning techniques are categorized into two categories: supervised or unsupervised learning. In supervised learning, we train the machine learning model using sample data with a labelled class or known as target variable. While unsupervised machine learning technique is designed to cluster the data into a few groups on the average value of distance between objects [10]. The following section covers some supervised machine learning classifiers and its applications.

2.3.1. Support vector machine

SVM is a supervised machine learning technique for classification problems. SVM uses a hyperplane which makes the decision to boundary between numerous classes. Hyperplane with the greatest margin is chosen to make predictions in the SVM algorithm. In general, SVM algorithm uses the input variables and attempts to find the ideal hyperplane that takes full advantage of the margin between two classes vectors [11]. Nagi *et al.* [12] used SVM classifier to detect non-technical loss for TNB in Malaysia using monthly electricity consumption in kWh, meter reading date and type, theft of electricity (TOE), credit worthiness rating (CWR), high risk customer (HRC), and irregularity report (IR). The predicted accuracy SVM was only 60%. Glauner [1] also used SVM for electricity data from Brazil and reported that his SVM model achieved 75% true positive rate (TPR) rate and false negative rate (FNR) at 25%. Jokar *et al.* [13] applied SVM to a large dataset of electricity usage (5000 customers of every 30 minutes for two years) and synthetically generated fraud data and reported SVM has achieved 94% prediction rate. Yap *et al.* [14] reported SVM for small dataset (n = 500) attained 83% accuracy, recommended that SVM performance is problem dependent, in which it based on the collected dataset, selected features and its split ratio.

2.3.2. Artificial neural network

ANN is brain-inspired simulation of the network of neurons which is intended to replicate the way humans learn so that the system able to learn things and make decisions in human manner. There are three layers in neural network which are input, hidden and output layer. The minimum hidden layer is one layer. Two or fewer hidden layers are sufficient with simple datasets. However, with complicated or complex datasets comprising computer vision or time series, extra layers can be useful [15]. Traditional ANN, the multi-layer perceptron (MLP) was used as a binary classifier for predicting electricity fraud and normally used for forecasting electricity consumption of time series data [4]. Jeyakumar and Devaraj [16] used electricity consumption data of 28 days recorded in 15 minutes time interval from Ireland. First, they applied k-means technique to group these customers into few clusters. Four clusters with the maximum number of customer profiles were selected. Since this dataset is single class (benign sample), synthetic fraud data was randomly generated from the selected clusters by multiplied average electricity reading per day with random value from -0.5 and 0.5. The ANN model reported has achieved 97% accuracy. ANN model was also applied to the same TNB dataset used by Sankari *et al.* [17] and the classifier accuracy was 92%. They recommended by identifying the most relevant features, we can improve the accuracy of the classifier.

2.3.3. Random forest

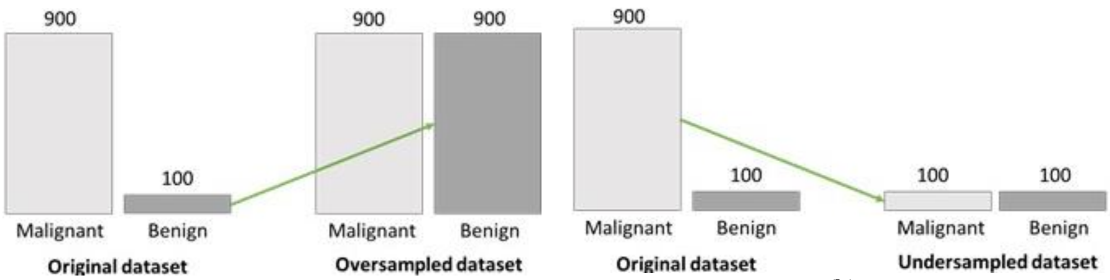
RF was developed by Breiman [18] and has become a popular machine learning model for classification and regression. RF uses ensemble classification tree where decision trees are developed using bootstrap samples of data and each tree used a random subset of the variable or features. The training algorithm for RF applies the bagging (bootstrap aggregating) method. Given a training set of sample size n, bagging repeatedly (B times) selects a random sample with replacement of the training sample and fits trees to the samples using a random subset of features sometime called “feature bagging”. The classification performance is then the average of the predictions (if Y is continuous) from the B samples or by taking most of the vote (if Y is categorical) for classification tree. Yang and Xu [19] used the minimum, maximum, mean, variance and medium of daily electricity consumption data plus the fraud data to develop the RF model. The accuracy of RF was at 98%. One large real-world dataset about 3.6 million customers consists of fraud inspection results, date and electricity consumption in kWh was studied by Glauner *et al.* [20] and RF achieved area under the curve (AUC) of 0.66 and 65% accuracy. Then in hologram, they visualized the prediction results of the customers and their neighborhoods so domain experts can review on which premises to perform the on-site inspection.

2.3.4. Extreme gradient-boosted tree

Gradient boosting is a machine learning technique for classification and regression problems. It is an ensemble model of weak prediction models, typically decision trees. For NTL prediction based on smart meter data, Buzau *et al.* [11] reported that XGBoost achieved AUC score of 0.91 when verified with real data from largest utility company in Spain. Buzau *et al.* [11] furthered asserts the major challenge of a machine learning model to accurately detect anomalies is the extremely imbalanced data between customer classes, where only 5% represents the fraudulent cases from the training dataset which can influence the process of learning, as the model would highly be biased to the majority class.

2.4. Imbalance datasets

Hordri *et al.* [21] acknowledged in the real dataset, the number of fraudulent is very small contrasted with the non-fraudulent class. Two basic sampling techniques are random over sampling (ROS) which duplicate oversample randomly for the minority class while RUS discards majority class to modify the class distribution. Oversampling may cause overfitting as it generates exact copies of the minority samples and creates synthetic data while undersampling may removes the potential meaningful majority samples [5], [21]. Figure 1 illustrated the concept of random oversampling and random undersampling. Figure 1(a) and Figure 1(b) illustrate the concept of random oversampling and oversampling respectively. In random oversampling, duplicates of samples of the minority class are generated to balance the sample, while in undersampling, samples are removed from the majority class to balance the sample.



(a) (b)

Figure 1. Random (a) oversampling and (b) undersampling for imbalanced dataset

The detected fraud behaviour (positive class) is very tiny fragment compared to excessively non-fraud data (negative class). On top of that, since the fraud samples are tagged manually by on-site inspections, human mistakes might have happened where the samples wrongly tagged [11] and they used RUS method for their study and confirms that by applying RUS, the AUC score improved significantly. On the other hand, few studies [12], [13], [16], [19], [22], [23] tackled this imbalanced data problem by generating synthetic data of fraud samples as shown in Figure 1. They have achieved more than 90% accuracy. Both studies by Glauner *et al.* [24] and Glauner [1] subsampled the imbalanced data into several proportions and made evaluations for all of them. Both oversampling and undersampling methods were proven effective in enhancing the classification accuracy for the electricity fraud imbalanced dataset. Imbalanced dataset is mostly addressed with two different approaches.

a) Data level approach such as oversampling the minority class or undersampling the majority class. SMOTE algorithm is frequently applied which repeatedly gains good results in imbalanced dataset classification [21].
b) Algorithm level approach, bagging and boosting, the most common methods are cost-sensitive learning and ensemble methods [25]. It is more profound that the model or algorithm accurately classifies the minority class rather than the majority class because the cost is higher if the model wrongly classifies the minority class [8].

2.5. Nature-inspired optimization algorithm

Most nature-inspired optimization algorithms are inspired from productive and successful characteristics of biological system. Some popular nature-inspired optimization algorithms are GWO [6], ant colony optimization [26], bat algorithm [27], cuckoo search [28], firefly algorithm [29], and particle swarm optimization [30]. Faris *et al*. [31] published a review in which these optimization algorithms have been widely used in various machine learning applications where these applications were categorized into four missions: feature selection, training neural networks, optimizing SVM and clustering applications.

2.5.1. Grey wolf optimizer

Various research applications have been using GWO due to its impressive performance such as in machine learning applications, engineering applications, wireless sensor network applications, environmental modelling applications, medical and bioinformatics application, and image processing applications. GWO is a meta-heuristics swarm intelligence technique proposed by Mirjalili *et al.* [6]. It was inspired from grey wolf behaviour which primarily impersonates grey wolf leadership ladder and their natural hunting method [6]. It has a special ability to target the exact balance between exploration and exploitation throughout the search leads to favourable convergence [31]. Jitkongchuen and Paireekreng [32] confirmed that GWO is one of the nature-inspired optimization algorithms that can solve the classification problems and handle imbalanced dataset including NP-hard problems. However, these nature-inspired optimization algorithms are problem specific, some algorithms would achieve better performance on other problems since the problems are varied, and even minor variances in the problem might favour specific approaches [33].

2.5.2. Particle swarm optimization

PSO algorithm was discovered by Ryenolds and Heppner where the algorithms were simulated by Kennedy and Eberhart [30]. PSO was inspired by social behaviour of bird flocking/roosting, animal herding, bacterial growth, and fish schooling. PSO is simple in concept, and it is for optimizing nonlinear functions [34]. PSO algorithm is similar with genetic algorithm and received extensive attention since it only uses few parameters, faster convergence, and other advantages [35]. PSO algorithm begin with initialization of the particles or swarm size population, followed by the initialization of inertia weight (*W*) and acceleration coefficient (*C1* and *C2*). Then initialize the minimum value (*V(min)*) and maximum value of velocity (*V(max)*) and minimum position (*Dmin*) and maximum value of position (*Dmax*), respectively. Next is the evaluation of *Pbest* and *Gbest* value for each particle and evaluates the new velocity value for each particle. Then, to updates the new position, *D(new)*. Finally, *Pbest(new)* and *Gbest(new)* are identified based on the fitness value. Iteration then continues to update the current velocity and position of each particle until it satisfies the stopping condition such as when the same maximum number of fitness value is reached. PSO has been applied in handling the imbalanced dataset for classification problem with numerous dataset sizes and has demonstrated its effectiveness in finding the best value with resulted an optimal balanced dataset [8], [35]. Li *et al.* [36] reported that PSO algorithm tackled the imbalanced dataset problem by increased the right volume of minority class which resulted better performance in their research.

2.6. Performance evaluations

The evaluation of the model should be performed on samples that are not used in model building, so that they keep an unbiased sense of model effectiveness [37]. Most classifiers performance are measures from computed confusion matrix. The true positive (TP) is the positive case which is correctly predicted while false positive (FP) is the negative case which is wrongly predicted as positive. True negative (TN) is the negative case which is correctly predicted while false negative (FN) is the positive case that is wrongly predicted as negative. Most classifiers performance measures are computed from the confusion matrix shown in Table 2. Several evaluations measures were considered: accuracy=(TP+TN)/(TP+FP+TN+FN), precision=TP/(TP+FP), sensitivity=TP/(TP+FN), specificity=TN/(TN+FP). There are also other measures to be considered such as false positive rate (FPR)=1-Specificity, false negative rate (FNR)=1-Sensitivity, F_1 score=2TP/(2TP+FP+FN), and geometric mean (G-Mean)=√(TPR×TNR). These measures are focus on optimizing the accuracy of each of the classes for binary classification. It is regularly used metric when dealing with imbalanced datasets [38].

Receiver operating characteristic (ROC) chart is a probability curve with a plot that visualize the performance of a binary classifier. It displays the predictive accuracy of a classifier model using sensitivity and specificity as a range cut-off of the model. In the ROC curve in Figure 2, when the curve is higher (red line-C model), it shows that the performance of the model is better as the AUC will be higher. The curve (light blue-model L) is nearer to the 45-degree red diagonal line light blue curve and indicate low AUC and classification performance. The AUC represents degree or measure of separability and ranges from 0 to 1. The higher the AUC or near to 1, the better the model as it shows the classifier can correctly distinguish between all the positive and the negative class. If, however, the AUC is 0, then the classifier would predict all negatives as positives and all positives as negatives.

Table 2. Confusion matrix

Electricity fraud	Electricity fraud predicted status	
Actual status	Fraud (1)	Non-fraud (0)
Fraud (1)	True positive (TP)	False negative (FN)
Non-fraud (0)	False positive (FP)	True negative (TN)

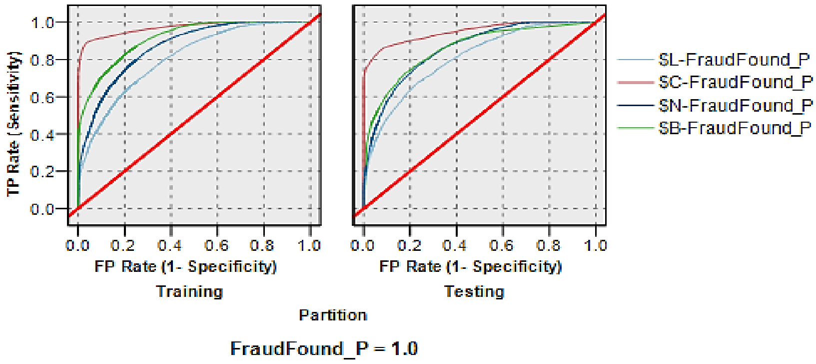


Figure 2. ROC chart

3. RESULTS AND DISCUSSION

This section presents the results, analysis, and the findings of this study. The focus of the study is to identify which machine learning classifier is effective in predicting electricity fraud based on customers consumption data and the most effective nature-inspired optimization algorithms, PSO and GWO in balancing the imbalance dataset to improve the classifiers performance. Moreover, the performance of all classifiers or four dataset will be presented to identify the most effective nature-inspired optimization algorithms.

3.1. Performance of machine learning classifiers

Predictive models for electricity fraud were developed using four classifiers which are ANN, SVM, XGBoost and RF. SVM was developed using radial basis function (RBF) kernel. For comparison purpose, each of this technique was applied to: i) the original imbalanced dataset ($Data_{ORI}$), ii) RUS (DataRUS) dataset, iii) PSO ($Data_{PSO}$) balanced dataset and iv) GWO ($Data_{GWO}$) balanced dataset. Table 3 shows the overall results for all four classifiers. The models were evaluated based on their accuracy, specificity, sensitivity, precision, F1 score rate, and AUC. Table 3 shows that, using the imbalanced dataset, although all ANN, SVM and XGBoost achieved high score on accuracy and specificity, their sensitivity and F1 score were very low. ANN model has the highest score for sensitivity (22.59%) and F1 score (31.35%) and the sensitivity for SVM was the lowest (0.88%). These results indicate that the classifiers were not able to perform well in predicting electricity fraud using the electricity consumption data which is imbalanced. The sensitivity for all four classifiers improved when data was balanced. The ANN classifier with balanced dataset using PSO achieved good performance with 94.38% accuracy, 92.81% sensitivity, 95.99% specificity, 95.97% precision and 94.36 F1 score, followed by ANN with GWO balanced dataset. The XGBoost classifier using $Data_{PSO}$ showed very good performance with 96.88% accuracy, 94.25% sensitivity, 99.58% specificity, 99.57% precision and 96.84 F1 score. RF using $Data_{PSO}$ has the highest performance with 96.98% accuracy, 94.87% sensitivity, 99.16% specificity, 99.14% precision and 96.96 F1 score. The performance for RF, XGBoost and ANN is also good under balanced $Data_{GWO}$.

Table 3. Classifiers performance (testing sample)

Model	Dataset	Accuracy	Sensitivity	Specificity	Precision	F1 Score
ANN	$Data_{ORI}$	90.55	22.59	97.73	51.24	31.35
	$Data_{RUS}$	64.52	50.31	79.11	71.22	58.97
	$Data_{PSO}$	**94.38**	**92.81**	**95.99**	**95.97**	**94.36**
	$Data_{GWO}$	85.43	83.37	87.55	87.31	85.29
SVM	$Data_{OR}$	90.50	0.88	99.98	80.00	1.74
	$Data_{RUS}$	59.00	23.61	95.36	83.94	36.86
	$Data_{PSO}$	**78.36**	**57.29**	**100**	**100**	**72.85**
	$Data_{GWO}$	77.73	56.06	100	100	71.84
XGBoost	$Data_{ORI}$	91.09	14.25	99.21	65.66	23.42
	$Data_{RUS}$	73.47	66.32	80.80	78.02	71.70
	$Data_{PSO}$	**96.88**	**94.25**	**99.58**	**99.57**	**96.84**
	$Data_{GWO}$	91.99	88.3	95.78	95.56	91.78
RF	$Data_{ORI}$	90.90	14.25	99.00	60.19	23.05
	$Data_{RUS}$	72.01	71.25	72.78	72.9	72.07
	$Data_{PSO}$	**96.98**	**94.87**	**99.16**	**99.14**	**96.96**
	$Data_{GWO}$	92.40	89.32	95.57	95.39	92.26

ROC chart in Figure 3 is used for comparison between the four classifiers for each four datasets. Figure 3(a) show the ROC chart for imbalance dataset, while Figures 3(b)-3(d) are the ROC for balanced dataset of RUS, PSO, and GWO respectively. For imbalanced dataset in Figure 3(a), the AUC score for RF is the highest with 0.808, quite similar with the score achieved by Zheng *et al.* [2] which was 0.801 using the same imbalanced dataset. However, their model, wide and deep convolutional neural networks attained 95.65% score on the mean average precision. XGBoost with AUC score 0.796 is the highest for RUS balanced dataset as shown in Figure 3(b). While for PSO and GWO balanced dataset illustrated by Figure 3(c) and Figure 3(d) respectively, showed that RF achieved the highest AUC scores with 0.989 and 0.967 respectively.

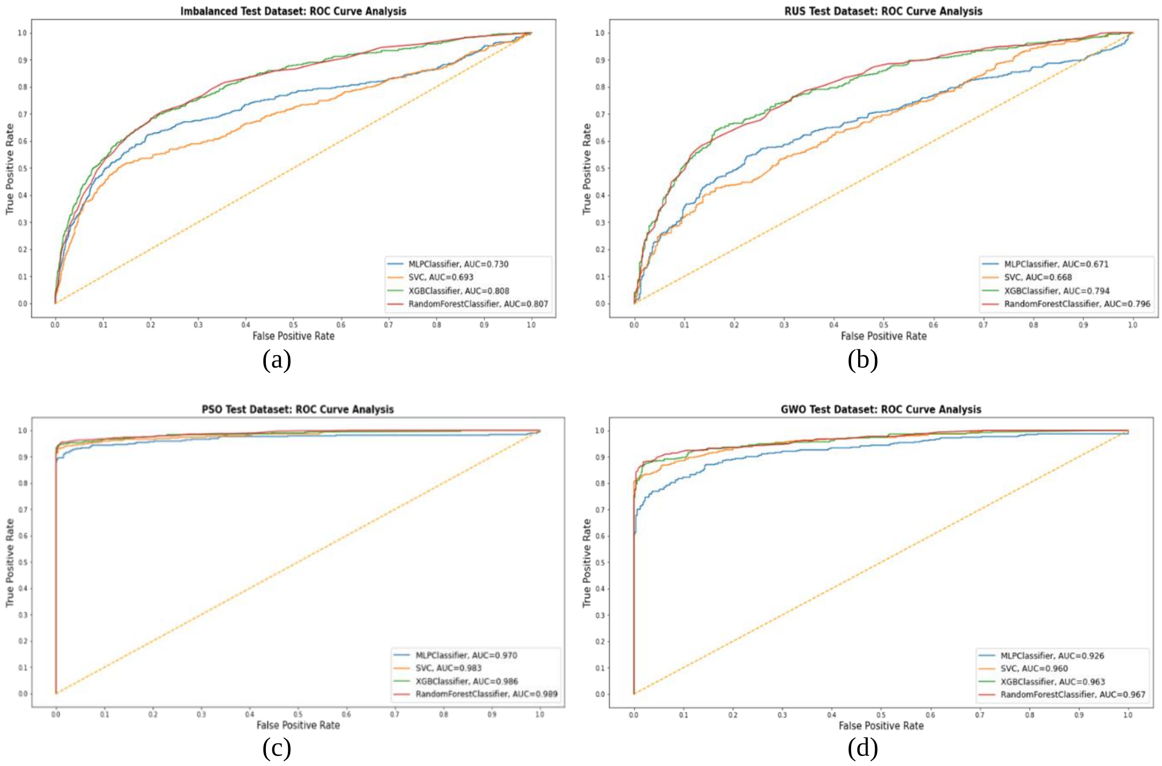


Figure 3. ROC chart for four classifiers: (a) Imbalanced dataset, (b) RUS balanced dataset, (c) PSO balanced dataset, and (d) GWO balanced dataset

3.2. Performance of hybrid machine learning classifiers

Figure 4 show the measurement performance of four machine learning classifiers computed using the confusion matrix. Figures 4(a)-4(d) show the performance of ANN, SVM, XGBoost, and RF classifiers for the four datasets respectively. Performance of all classifiers improved tremendously when balancing using PSO and GWO. RF performed well for balanced datasets. Performance of SVM improved when hybrid with PSO or GWO. Three of the classifiers, ANN, XGBoost, and RF have achieved more than 92% in all their classification performance measurements. SVM has the lowest performance compared to ANN, RF and XGBoost.

Table 4 displays the results of four classifiers hybrid with PSO balanced dataset. The results show that balancing sample using PSO improved the four classifiers performance tremendously for all with high accuracy (ANN=94.38%, SVM=78.36%, XGBoost=96.88%, RF=96.98%), specificity (ANN=95.99%, SVM=100%, XGBoost=99.58%, RF=99.16%) and precision (ANN=95.97%, SVM=100%, XGBoost=99.57%, RF=99.14%) for ANN, XGBoost and RF. Three of the classifiers (ANN, XGBoost and RF) achieved more than 92% in all their classification performance measurements. Although SVM achieved high score for specificity, precision and AUC, the accuracy, sensitivity and F1 score were the lowest among the four models.

Table 5 displays the results for the four classifiers hybrid with GWO balanced dataset. The three classifiers, ANN, XGBoost and RF achieved slightly lower performance if compared with PSO. However, SVM still suffers low score in the sensitivity for PSO balanced data (ANN=92.81%, SVM=57.29%, XGBoost=94.25%, RF=94.87%) and GWO balanced data (ANN=83.37%, SVM=56.06%, XGBoost=88.33%, RF = 89.32%).

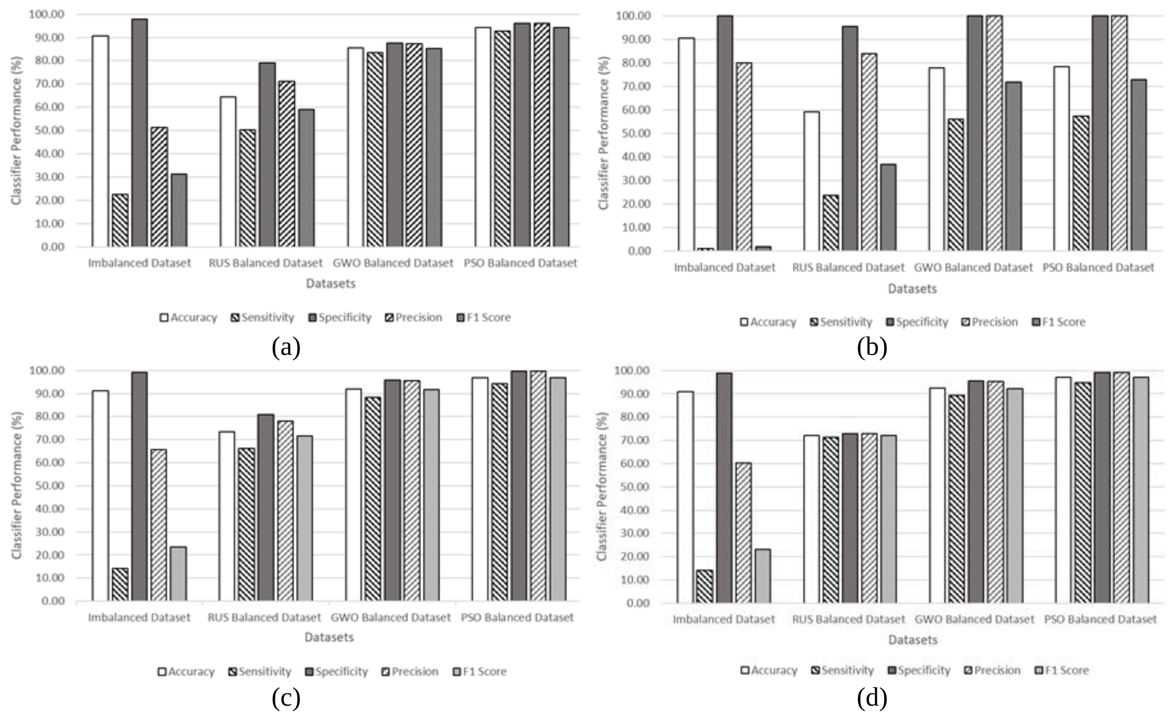


Figure 4. Machine learning classifier performance (testing samples) (a) ANN, (b) SVM, (c) XGBoost, and (d) RF

Table 4. Comparison of machine learning classifier hybrid PSO balanced dataset

Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score	AUC
ANN	94.38	92.81	95.99	95.97	94.36	0.970
SVM	78.36	57.29	100.00	100.00	72.85	0.983
XGBoost	96.88	94.25	99.58	99.57	96.84	0.986
RF	96.98	94.87	99.16	99.14	96.96	0.989

Table 5. Comparison of machine learning classifier hybrid GWO balanced dataset

Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score	AUC
ANN	85.43	83.37	87.55	87.31	85.29	0.926
SVM	77.73	56.06	100	100	71.84	0.960
XGBoost	91.99	88.30	95.78	95.56	91.78	0.963
RF	92.4	89.32	95.57	95.39	92.26	0.967

Based on the evaluation of the performance results and ROC charts, RF classifier hybrid with PSO has outperformed other classifiers with highest score in its accuracy, sensitivity, specificity, precision, F1 score and AUC for balanced datasets. It is closely followed by XGBoost classifier hybrid with PSO. Thus, combination of RF classifier hybrid with PSO is selected as the best hybrid method for electricity fraud prediction. These results also showed that nature-inspired algorithms especially PSO is an effective algorithm to perform undersampling technique to address imbalance class problem.

4. CONCLUSION

This study shows that nature-inspired algorithms especially PSO can optimize the performance of machine leaning classifiers for electricity fraud prediction especially on imbalanced datasets problem. This study introduced the hybrid method by combining the nature-inspired optimization algorithms together with machine learning classifiers. PSO and GWO algorithms to address the imbalanced problem by undersampling the majority class. The nature-inspired algorithms involve the mathematical formulations to provide optimal solution or fitness value to undersample the majority class which have high diversity between the samples. The performance of the hybrid method was evaluated and has shown tremendous improvement for all four classifiers. When data is balanced using PSO or GWO, three classifiers (ANN, XGBoost and RF) achieved very high-performance score in predicting the electricity fraud, not only for sensitivity and F1 score, but for all performance measures. Future research can explore the capability of nature-inspired optimization algorithms and the improvement of the classifier in other domains as well such as business, finance, medical and healthcare data where the imbalanced class exists.

ACKNOWLEDGEMENTS

The authors wish to express their thanks to College of Computing. Informatics and Mathematics and Research Management Centre (RMC), Universiti Teknologi MARA, Shah Alam for supporting the publication of this paper.

REFERENCES

[1] P. Glauner, "Artificial intelligence for the detection of electricity theft and irregular power usage in emerging markets," Ph.D. dissertation, University of Luxembourg, Luxembourg, 2019.

[2] Z. Zheng, Y. Yang, X. Niu, H. N. Dai, and Y. Zhou, "Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids," *IEEE Transactions on Industrial Informatics*, vol. 14, no. 4, pp. 1606–1615, =2018, doi: 10.1109/TII.2017.2785963.

[3] "TNB to go after account holders to recover millions lost to power theft, new straights times," NST, 14 August 2019, [Online]. Available: https://www.nst.com.my/business/2019/08/512831/tnb-go-after-account-holders-recover-millions-lost-power-theft.

[4] G. M. Messinis and N. D. Hatziargyriou, "Review of non-technical loss detection methods," *Electric Power Systems Research*, vol. 158, pp. 250–266, May 2018, doi: 10.1016/j.epsr.2018.01.005.

[5] B. W. Yap, K. A. Rani, H. A. Abd Rahman, S. Fong, Z. Khairudin, and N. N. Abdullah, "An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets," in *Lecture Notes in Electrical Engineering*, vol. 285 LNEE, Springer Singapore, 2014, pp. 13–22, doi: 10.1007/978-981-4585-18-7-2.

[6] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, Mar. 2014, doi: 10.1016/j.advengsoft.2013.12.007.

[7] Y. Dong and C. Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus*, vol. 2, no. 1, pp. 1–17, May 2013, doi: 10.1186/2193-1801-2-222.

[8] J. Li *et al.*, "Adaptive swarm balancing algorithms for rare-event prediction in imbalanced healthcare data," *PLoS ONE*, vol. 12, no. 7, p. e0180830, Jul. 2017, doi: 10.1371/journal.pone.0180830.

[9] A. Haya, "Handling class imbalance using swarm intelligence techniques, hybrid data and algorithmic level solutions," Ph.D. dissertation, Goldsmiths, University of London, 2019).

[10] Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: a review," *Energy Informatics*, vol. 1, no. 1, Aug. 2018, doi: 10.1186/s42162-018-0007-5.

[11] M. M. Buzau, J. Tejedor-Aguilera, P. Cruz-Romero, and A. Gomez-Exposito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661–2670, 2019, doi: 10.1109/TSG.2018.2807925.

[12] J. Nagi, K. S. Yap, S. K. Tiong, S. K. Ahmed, and M. Mohamad, "Nontechnical loss detection for metered customers in power utility using support vector machines," *IEEE Transactions on Power Delivery*, vol. 25, no. 2, pp. 1162–1171, Apr. 2010, doi: 10.1109/TPWRD.2009.2030890.

[13] P. Jokar, N. Arianpoo, and V. C. M. Leung, "Electricity theft detection in AMI using customers' consumption patterns," *IEEE Transactions on Smart Grid*, vol. 7, no. 1, pp. 216–226, Jan. 2016, doi: 10.1109/TSG.2015.2425222.

[14] K. Yap, S. Tiong, J. Nagi, J. Koh, and F. Nagi, "Comparison of supervised learning techniques for non-technical loss detection in power utility," *International Review on Computers and Software*, vol. 7, no. 2, pp. 626–636, 2012.

[15] J. Heaton, "The number of hidden layers," *Heaton Research*, June 2017 [Online]. Available: https://www.heatonresearch.com/2017/06/01/hidden-layers.html,.

[16] J. Jeyakumar and D. Devaraj, "Machine learning algorithm for efficient power theft detection using smart meter data," *International Journal of Engineering and Technology*, vol. 7, pp. 900–904, 2018.

[17] E. S. Sankari, R. Rajesh, and P. Matheswaran, "Detection of non-technical loss in power utilities using data mining techniques," *International Journal for Innovative Research in Science and Technology*, vol. 1, no. 9, pp. 97–101, 2015.

[18] L. Breiman, "Random forests," *Random forests*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[19] Y. Yang and G. Xu, "Theft of electricity detection based on supervised learning," *Yau Science Award (Computer Science)*, 2017.

[20] P. Glauner *et al.*, "Identifying irregular power usage by turning predictions into holographic spatial visualizations," in *IEEE International Conference on Data Mining Workshops, ICDMW*, Nov. 2017, vol. 2017-Novem, pp. 258–265, doi: 10.1109/ICDMW.2017.40.

[21] N. F. Hordri, S. S. Yuhaniz, N. F. M. Azmi, and S. M. Shamsuddin, "Handling class imbalance in credit card fraud using resampling methods," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 11, pp. 390–396, 2018, doi: 10.14569/ijacsa.2018.091155.

[22] A. Aligholian, M. Farajollahi, and H. Mohsenian-Rad, "Unsupervised learning for online abnormality detection in smart meter data," in *IEEE Power and Energy Society General Meeting*, Aug. 2019, vol. 2019-Augus, doi: 10.1109/PESGM40551.2019.8973564.

[23] R. Sowndarya and P. Latha, "An artificial intelligent algorithm for electricity theft detection in AMI," *International Journal of Engineering Science and Computing*, vol. 7, no. 3, pp. 5222–5227, 2017.

[24] P. Glauner *et al.*, "Large-scale detection of non-technical losses in imbalanced data sets," *2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, pp. 1-5, 2016, doi: 10.1109/ISGT.2016.7781159.

[25] V. Karia, W. Zhang, A. Naeim, and R. Ramezani, "GenSample: a genetic algorithm for oversampling in imbalanced datasets," *arXiv preprint arXiv:1910.10806*, 2019.

[26] M. Dorigo and G. Di Caro, "Ant colony optimization: A new meta-heuristic," in *Proceedings of the 1999 Congress on Evolutionary Computation, CEC 1999*, 1999, vol. 2, pp. 1470–1477, doi: 10.1109/CEC.1999.782657.

[27] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," *In Nature inspired cooperative strategies for optimization (NICSO 2010)*, pp. 65–74, 2010.

[28] X. S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *2009 World Congress on Nature and Biologically Inspired Computing, NABIC 2009 - Proceedings*, 2009, pp. 210–214, doi: 10.1109/NABIC.2009.5393690.

[29] Y.-S. Yang, "Firefly algorithm," *Nature-inspired metaheuristic algorithms*, vol. 20, pp. 81–89, 2010.

[30] J. Kennedy and R. Eberhart, "Particle swarm optimization," *Proceedings of ICNN'95 - International Conference on Neural Networks*, 1995, pp. 1942-1948 vol.4, doi: 10.1109/ICNN.1995.488968.
[31] H. Faris, I. Aljarah, M. A. Al-Betar, and S. Mirjalili, "Grey wolf optimizer: a review of recent variants and applications," *Neural Computing and Applications*, vol. 30, no. 2, pp. 413–435, Nov. 2018, doi: 10.1007/s00521-017-3272-5.
[32] D. Jitkongchuen and W. Paireekreng, "The Optimized Function Selection Using Wolf Algorithm for Classification," *Proceedings of International Conference on Artificial Life and Robotics*, vol. 22, pp. 563–566, Jan. 2017, doi: 10.5954/icarob.2017.gs10-1.
[33] M. A. Lones, "Mitigating metaphors: a comprehensible guide to recent nature-inspired algorithms," *SN Computer Science*, vol. 1, no. 1, Nov. 2020, doi: 10.1007/s42979-019-0050-8.
[34] A. Tharwat, A. Hassanien, T. Gaber, and B. E. Elnaghi, "Particle swarm optimization - a tutorial," *Handbook of research on machine learning innovations and trends. IGI Global*, pp. 614–635, 2017.
[35] J. Li, S. Fong, and Y. Zhuang, "Optimizing SMOTE by metaheuristics with neural network and decision tree," in *Proceedings - 2015 3rd International Symposium on Computational and Business Intelligence, ISCBI 2015*, Dec. 2016, pp. 26–32, doi: 10.1109/ISCBI.2015.12.
[36] J. Li, S. Fong, S. Mohammed, and J. Fiaidhi, "Improving the classification performance of biological imbalanced datasets by swarm optimization algorithms," *Journal of Supercomputing*, vol. 72, no. 10, pp. 3708–3728, Nov. 2016, doi: 10.1007/s11227-015-1541-6.
[37] M. Kuhn and K. Johnson, "Applied predictive modeling," *Applied predictive modeling*, vol. 26, no. 13, 2013.
[38] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Computing Surveys*, vol. 49, no. 2, pp. 1–50, Aug. 2016, doi: 10.1145/2907070.

BIOGRAPHIES OF AUTHORS

Ami Shamril Kamaruddin received his M.Sc. degree in data science from Universiti Teknologi MARA, Malaysia. He is currently working with one of the banks in Malaysia managing in-house and cloud applications by adapting DevOps practices and principles. He can be contacted at email: ashamril@gmail.com.

Mohd Fikri Hadrawi received his B.Sc. degree in statistics from Universiti Teknologi MARA, Malaysia. He then obtained his M.Sc. degree in statistics from the University of Malaya, Malaysia. He is currently a lecturer at School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA (UiTM). His research interests include computational statistics, probability and statistics, distribution theory, statistical modeling, and experimental research and design. He can be contacted at email: fikri1814@uitm.edu.my.

Yap Bee Wah graduated with a B.Sc. (Hons) (Mathematics Education) from Universiti Sains Malaysia. She then obtained her Master of Statistics degree from University of California, Riverside, USA, and her Ph.D. (Statistics) from University of Malaya, Malaysia. She has more than 30 years of service at Universiti Teknologi MARA, Malaysia, and recently joined UNITAR International University as Director of the Research and Consultancy Centre. Her research interests are in big data analytics and data science, computational statistics, and multivariate data analysis. Her research works are in the healthcare, education, environment, and business domains. She has published more than 100 papers in Scopus-indexed journals and proceedings. She can be contacted at email: bee.wah@unitar.my.

Sharifah Aliman received B.Sc. in Computer Science from Midwestern State University, Texas, USA, M.Sc. (Information Technology), Universiti Putra Malaysia, and Ph.D. in Information Technology and Quantitative Sciences, Universiti Teknologi MARA. She previously held faculty positions as coordinator program for Diploma in Computer Science (2007-2009), coordinator program for Master of Computer Science and Master of Data Science (2018-2019), coordinator program by research (2019-2021) and Head of Postgraduate Studies at Faculty of Computer and Mathematical Sciences (2020-2021). Currently, Head of Academic Management at the Office of The Deputy Vice Chancellor (Academic and International) and a senior lecturer at Universiti Teknologi MARA, Shah Alam, Malaysia. Her research interests include social media analytics, data analytics and scientific visualization. She can be contacted at email: sharifahali@uitm.edu.my.