Performance Evaluation Using Machine Learning: Detecting Non-technical Losses in Smart Grid

P. Abhinayaa, R. Ezhilarasie, and A. Umamakeswari

Abstract In the generation and distribution of electricity in the power grid, there is a chance for the occurrence of non-technical loss while transmitting by various means. One of the most concern needed loss varieties is electricity theft. The theft occurrence may cause significant loss and harm to the power grid and also to the economy by leading to unprofitable accounts for the power supply companies. Regular inspection on irregular consumption of power is inefficient and very time consuming. Utilizing machine learning in this theft detection system helps to prevent huge losses. In this paper, various machine learning models are employed to state the better performing model for the given data. By employing the various techniques of machine learning, an effective model for theft detection can be obtained and the problem associated with non-technical loss especially theft detection can be monitored and controlled.

Keywords Electricity theft detection · Non-technical losses · Machine learning algorithms

1 Introduction

In the world of modern life, electricity has turned to be a very mandatory thing since without it every work to be done seems to be impossible. Such kind of mandated life element needs to be conserved and should use it efficiently. The effective usage of electricity is turned to be an unlearned art for both the electricity providers and consumers. On seeing from the side of electricity providers, the major thing that should be considered and regulate is electricity loss [1] occurring while generation and distribution of it. As in [2], the electricity losses can be said as the occurrence of technical loss and non-technical loss. The technical losses are due to machinery problems occurring during the generation of electricity. In the case of non-technical losses [3], loss occurs when there is a chance for incorrect meter reading, improper meter installation and theft [4]. Among them, theft is a very serious issue facing everywhere

P. Abhinayaa · R. Ezhilarasie (✉) · A. Umamakeswari
SASTRA Deemed University, Thanjavur, TN 613401, India

S. Rawat et al. (eds.), *Proceedings of First International Conference on Computational Electronics for Wireless Communications*, Lecture Notes in Networks and Systems 329,
https://doi.org/10.1007/978-981-16-6246-1_19

by the power-producing companies. The electricity theft makes the economy of the government to get degrade as the customer won't pay the bill that they have actually consumed, which causes the revenue loss that should return as a profit. The impact of loss is not limited to the degradation of power quality [5]. The theft causes load imbalance in the grid that makes the electricity provider not meet the demand needs of that region. The increase in the demand may result in voltage drop, transformer overload, etc., that affect the corresponding power line and in the worst situation, put the life of the public in danger. Hence, the theft in electricity should get detected and vanish.

The theft occurrence can be monitored by the conventional method that is in-person inspection and verification on the customer and their usage by the corresponding authorities but it is very time consuming and can be manipulated by the corrupted people. As an alternate method, the installation of smart meters [6, 7] and implementation of machine learning algorithms [8, 9] come into play.

Employing a machine learning algorithm for this process is a very useful and simple way to monitor and detect theft occurrence on power and pattern of consumption rates. Employing the machine learning techniques in the smart grid which comprises various power producing units and sectors is very useful to protect the grid connectivity from damage caused by any fluctuation due to load imbalances. With the implementation of this detection system, the grid can be assured for theft prone as it categorizes the anonymous activity from the usual behavior while reviewing the consumption pattern of the smart grid. Here various algorithms were discussed to obtain the optimized better performing model for the given data. The historical data were used to frame the various machine learning model. The dataset was released by the State Grid Corporation of China (SGCC) [10] from that 1035 days were taken into account for the process of framing the model of machine learning.

This paper proposes a comparative analysis of different machine learning model solutions for energy theft detection. In Sect. 2, it compares the works and theory related to proposing work that exists. Section 3 presents the stages that the model undergoes while performing the theft detection process. In Sect. 4, the obtained results are discussed and analyzed with various performance metrics. Section 5 concludes the result and presents the outcome of the proposed work.

2 Related Works

This section presents the existing works that are related to theft and fraud occurrence detection in power systems for both the traditional and smart grid networks.

Different approaches were implemented to realize the rate of energy production, monitoring and control and forecasting of energy production, distribution, energy loss either by means of technical loss or non-technical loss. However, the detection of theft occurrence in the smart grid plays a vital role in the reliability of the consumer. As the detection process needs accurate pointing of the fraudulent customer. Using smart meter data on the advanced metering infrastructure (AMI) in the smart grid is helpful to detect the occurrence of electricity theft [11]. On the other hand, the AMI is prone to other techniques of theft attacks [12] especially by means of cyber-attacks and using digital tools. In order to solve these kinds of issues arising, many techniques were put forward to overcome the drawbacks resulted from various means. The state-based detection [13] model is based on the combination of distribution transformer and wireless sensors [14]. This model is dependent on the real-time data acquisitions of physically measured that are unattainable on some occasions and also opens a door to cyber-attacks on it where the data can be altered illegally. As in [15, 16] game-based detection model, support the process of theft detection by establishing a game between the power utility and the theft from which the normal and abnormal, that is, fraudulent and non-fraudulent characteristics can be obtained from the game equilibrium. Using a game-based detection model, it is possible to achieve a low cost and reasonable result of theft reduction but the establishment of utility function for all players is a challenging task.

From [17], it is agreeable that machine learning should be deployed to identify the possible occurrence of fraudulent behavior, however after that the physical inspection should take place. And also, it insisted on the necessity of thinking wider social, economic, and legal considerations should not be neglected as a way of reducing the loss. The non-technical loss also appears by means of cyber-attacks on the distribution network itself, using the preventive methods provided in [18] it can be detected. According to the survey of [19], most of the solutions of detection techniques of electricity theft lie in two wide circles as an expert system and machine learning model. The expert system seeks human experts to solve the problem by following the regulations with user-defined rules. However, such a system is very time consuming, and most importantly there is a chance of biases in judgments while the decision-making process. Machine learning solutions are emerging as a popular alternative [20] and they will perform effectively with the support of the availability of large quantities of data that are obtained from smart meters. For machine learning techniques, it is easier to learn defined patterns from historical data and it reduces the need for being explicitly programmed to it.

The algorithms that help for machine learning solutions can be segregated into clustering (unsupervised) and classification (supervised) models. [21] supports the theft detection solution by using the principal component analysis, [22] uses K-means clustering technique, [23] uses a C-fuzzy technique, but this method has a drawback in terms of accuracy though fuzzy gives good accuracy there are still the chances that the training set fuzzy clusters may not yield an accurate load details. Although clustering-based machine learning detection solutions are remarkable, their scaled performances were still not far enough to reach the real-time implementation. Hence classification techniques come into play. There are different algorithms involved, [24] details the theft detection solution based on the support vector machine technique where the desired detection hit rate of 60% was achieved and this rate is improved in [25] by 70%. Similarly, [26] supports the K-nearest neighbor. Algorithms based on supervised learning methodologies produce a good result for real-time needs comparatively.

3 Methodology Description

In this work, five different algorithms are taken into consideration to frame the solution to the electricity theft detection system. They are logistic regression (LR) [27], Support vector machine (SVM), Naive Bayes (NB) [28], Decision tree (DT) [29], and Random forest (RF) [30]. The data set is collected from the state grid corporation of china. The data consist of details of the electricity consumption value of 42,372 customer details for a period of 1035 days from 01 January 2014 to 09 September 2016. Using this dataset, the analysis of different algorithms is made. The theft detection solution undergoes the stages like data preprocessing, generation of train, test and validation set, building of machine learning algorithm, and performance analysis of the different techniques that were taken into consideration. The overall flow of the process is described in Fig. 1.

Data preprocessing. The data that are provided as an input to any machine learning model should be preprocessed and verified to avoid the confusion that occurs at the algorithm as it causes more generalization on the learning model. The dataset taken is primarily preprocessed by removing the noise/outliers present in the dataset. Then, the outlier-removed data set will get checked for presence for any null values. It can also be said as checking for missing values which is referred to as data imputation.

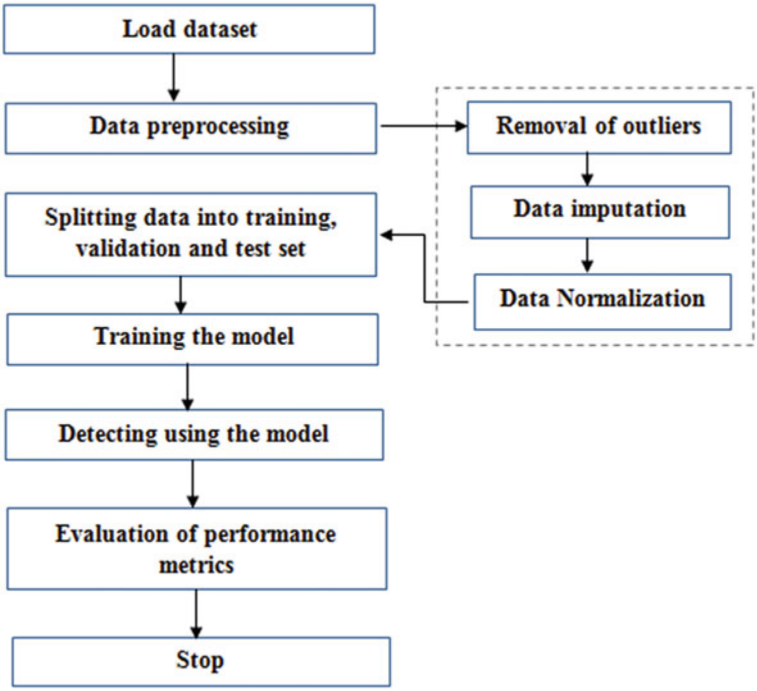
Load dataset
Data preprocessing
Removal of outliers
Data imputation
Data Normalization
Splitting data into training, validation and test set
Training the model
Detecting using the model
Evaluation of performance metrics
Stop

Fig. 1 Overall flow of the proposed work

Finally, after completing the aforesaid stages, the data present in the dataset will get normalized by min–max techniques for a range of (0, 1). The normalization should be done to avoid the suffering of a machine learning model with a diverse range of data.

Generation of train, test, and validation set. The dataset consists of details of the behavior of customer of 42,372 where the count of details of an honest user is more than the dishonest user. It causes the dataset to be imbalanced [31]. In order to overcome the imbalanced dataset, the oversampling technique, SMOTE is applied [32]. Hence, the minority count of fraudulent customer details got oversampled and increased the count equal to the non-fraudulent customer details. After applying the SMOTE algorithm, the dataset has to get split for the process of training, testing, and validation of the dataset.

Building machine learning models. Accordingly, the segregated set of data will be used for training the model. The different machine learning models will process the data given to yield results. Since the learning models use the hyper-parameters, the parameters got optimized and selected by using the grid search method [33]. The tuned parameter is given to the model. For example, the parameter like maximum depth, number of estimators for the random forest is selected based on the grid search method.

4 Result and Analysis

The implementation was done with the help of Python 3.6. Experiments are conducted with the support of Intel Core i3 with 4.0 GB of RAM on a standard PC in the virtual environment called Google Collaboratory.

Evaluation of the proposed work is an important criterion to be followed as it describes the nature of the model and how well the model achieves the objective of the work. However, it is not enough to fully judge the model but helps to understand the performance level of the model. The evaluation metrics include different types like classification accuracy, logarithmic loss, confusion matrix [34], area under curve, f1 score, mean absolute error, and mean squared error. Here, the mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE) are used to evaluate the model.

Different machine learning models are evaluated to find an accurate prediction scheme.

4.1 Logistic Regression (LR)

Logistic regression is used to predict values within a continuous range rather than trying to classify them into categories. In LR, parameter *C* is considered and given to grid search to pick the value that yields higher accuracy when applied. For that, the value for C ranges from $1e{-}7$ to $1e0$. With the tuned parameter value, LR produces 0.72 as an accuracy value. Also, the MAE, MSE, and RMSE are computed as 0.27, 0.26, and 0.52, respectively, and can graphically view this in Fig. 5.

4.2 Decision Tree (DT)

In the decision tree, while calculating the target value of a model, the predictive model uses binary rules and in this model, each individual tree has branches, nodes, and leaves. Parameters like max_depth, min_sample_split and criterion are considered for grid search method with the values 1–8 for max_depth, 2–4 for min_sample_split and Gini and entropy for criterion. With those values, it had scored the accuracy value as 0.73 which is much near to the previously discussed model logistic regression. As shown in Fig. 5, the error calculated for this model is 0.26, 0.26, and 0.51 for MAE, MSE, and RMSE, respectively.

4.3 Random Forest (RF)

It is a specialized decision tree where multiple decision trees got integrated to achieve better performance. It helps to maintain the distinctive control of overfitting than implementing with a single decision tree. The RF classifier can handle data that are with high-dimensionality while maintaining computational efficiency higher. The parameters like max_depth, max_features, min_sample_leaf, min_sample_split, n_estimators are taken into consideration for the tuning process. It has produced the result when computed as 91.96% that is 0.92 as accuracy value with the MAE, MSE, and RMSE as 0.08, 0.08, and 0.2 which is shown in Table 1 and Fig. 5.

4.4 Naive Bayes (NB)

Mostly due to the NB's oversimplified assumptions, this classifier works in a much better way in many complex real-world situations. Here, the model is reported the 60% as accuracy with the error rate of 0.39 as MAE, 0.39 as MSE, and 0.63 as RMSE when the parameter var_smoothing is tuned. This can be visualized in Fig. 5.

Table 1 Experimental parameters used in the discussed algorithm

Algorithms	Parameters	Values
LR	Inverse regularization strength	1.0
	Penalty	l2
	n_jobs	−1
NB	vaar_smoothing	$1e{-}9$
DT	Criterion	entropy
	max_depth	8
	min_samples_split	3
RF	max_depth	100
	max_feature	3
	min_sample_leaf	3
	min_sample_split	12
	n_estimators	100
SVM	C	0.7
	gamma	0.1
	Kernel	rbf

4.5 Support Vector Machine (SVM)

With the help of hyper-parameters like C, gamma, and kernel the support vector model performs the desired work and yields the output with the accuracy of 71.9% along with the error of 0.28 in MAE and 0.52 in RMSE where the comparison between all the model's error value can be seen in Table 2 and Fig. 5.

Overall, it is found that the random forest algorithm outperformed the various kinds of machine learning algorithms. Also, the AUC value obtained from the ROC plot for the model random forest is 0.98, this can be seen in Fig. 2. The different values obtained during the evaluation of different algorithms are plotted in the graph as shown in Figs. 3 and 4 where the precision, recall, f1 score are taken into consideration as some of the performance metrics for both non-fraudulent (Class 0) and fraudulent customers (Class 1), respectively. From the before-mentioned Figs. 3 and 4, it is clearly interpretable that the model random forest outperformed the remaining model by means of the three performance metrics that are taken into account. Similarly, from the graph of Fig. 5, it is crystal clear that the loss occurring for the same model is comparatively low.

Table 2 Computed accuracy of different models in percentage

Accuracy for different models				
LR	DT	NB	RF	SVM
72.02	73.33	60.11	92	71.9

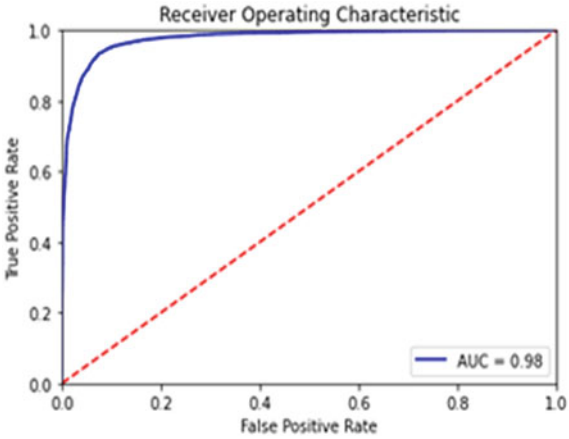

Fig. 2 ROC of random forest which performed better in the overall analysis

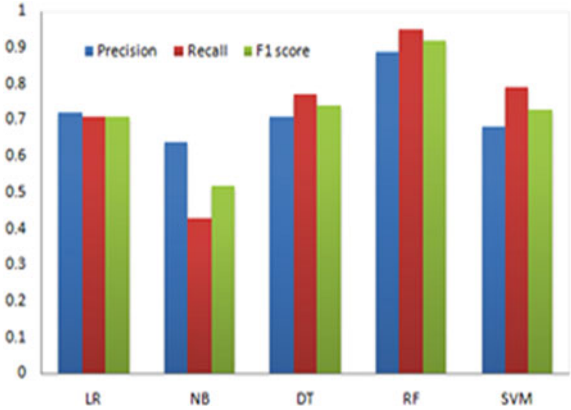

Fig. 3 Performance metrics for different algorithms for test data (non-fraudulent)

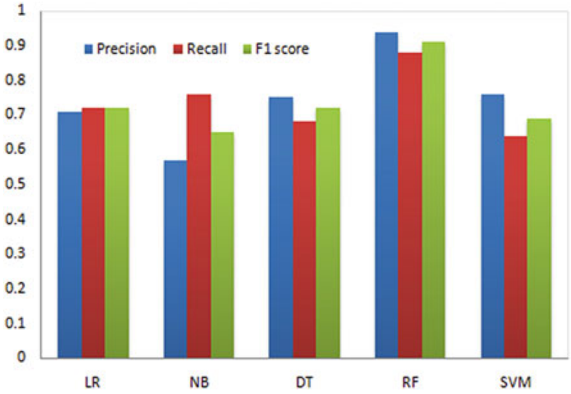

Fig. 4 Performance metrics for different algorithms for test data (fraudulent)

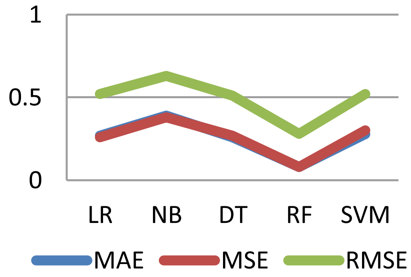

Fig. 5 Graph of error rate MAE, MSE, and RMSE obtained for the test

5 Conclusion

In the work described, different types of machine learning algorithms are proposed for the process of theft detection happening in the smart grid. Different conventional as well as modern methods of machine learning methods used for theft detection have been discussed and analyzed with their merits and demerits. These methods use historical power consumption data for detection. There are many types of evaluation criteria for checking the accuracy and error/loss of these models and the same has been identified and used to evaluate the test data in this paper. Based on the results obtained, random forest performed well, as it progresses based on bagging by considering the needed set of features rather than all the features and another advantage of RF is that little preprocessing and can be parallelizable.

Acknowledgements The authors wish to express their sincere thanks to the Department of Science & Technology, New Delhi, India (Project-ID: SR/FST/ETI-371/2014), SASTRA Deemed University, Thanjavur, India, for extending the infrastructural support to carry out this work.

References

1. L. Weijun et al., Research on transmission line power losses effected by harmonics, in *China International Conference on Electricity Distribution (CICED)*, Xi'an (2016), pp. 1–3
2. A. Al-Hinai, A. Al-Badi, E. Feilat, M. Albadi, H. Al-Nassri, A. Al-Busaidi, Energy losses in power system—practical case study (2012)
3. P. Chandel, T. Thakur, B.A. Sawle, R. Sharma, Power theft: major cause of non-technical losses in Indian distribution sector, in *IEEE 7th Power India International Conference (PIICON)*, Bikaner (2016), pp. 1–6
4. M. Golden, B. Min, Corruption and theft of electricity in an Indian state (2011)
5. L.G. Arango, E. Deccache, B.D. Bonatto, H. Arango, P.F. Ribeiro, P.M. Silveira, Impact of electricity theft on power quality, in *17th International Conference on Harmonics and Quality of Power (ICHQP)*, Belo Horizonte (2016), pp. 557–562
6. S. Sahoo, D. Nikovski, T. Muso, K. Tsuru, Electricity theft detection using smart meter data, in *IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, Washington, DC (2015), pp. 1–5
7. S.S.S.R. Depuru, L. Wang, V. Devabhaktuni, Electricity theft: overview, issues, prevention and a smart meter-based approach to control theft. Energy Policy 1007–1015 (2011)

8. J. Jeyaranjani, D. Devaraj, Machine learning algorithm for efficient power theft detection using smart meter data. Int. J. Eng. Technol. 900–904 (2018)
9. N. Dahringer, Electricity theft detection using machine learning (2017)
10. X. Yi-chong, The state grid corporation of China, in *The Political Economy of State-Owned Enterprises in China and India. International Political Economy Series*, ed. by X. Yi-chong (Palgrave Macmillan, London, 2012)
11. S.K. Singh, R. Bose, A. Joshi, Energy theft detection in advanced metering infrastructure, in *IEEE 4th World Forum on Internet of Things (WF-IoT)*, Singapore (2018), pp. 529–534
12. R. Jiang, R. Lu, Y. Wang, J. Luo, C. Shen, X. Shen, Energy-theft detection issues for advanced metering infrastructure in smart grid. Tsinghua Sci. Technol. 105–120 (2014)
13. S.-C. Huang, Y.-L. Lo, C.-N. Lu, Non-technical loss detection using state estimation and analysis of variance. IEEE Trans. Power Syst. **28**(3), 2959–2966 (2013)
14. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, Wireless sensor networks: a survey. Comput. Netw. **38**(4), 393–422 (2002)
15. S. Amin, G.A. Schwartz, A.A. Cardenas, S.S. Sastry, Game-theoretic models of electricity theft detection in smart utility networks: providing new capabilities with advanced Journal of Electrical and Computer Engineering metering infrastructure. IEEE Control Syst. Mag. **35**(1), 66–81 (2015)
16. D. Yao, M. Wen, X. Liang, Z. Fu, K. Zhang, B. Yang, Energy theft detection with energy privacy preservation in the smart grid. IEEE Internet Things J. **6**(5), 7659–7669 (2019)
17. V. Krishna, G.A. Weaver, W.H. Sanders, Non-technical losses in the 21st century: cause, economic effects, detection and perspectives. Technical Report (University of Luxembourg, 2018)
18. J. Leite, J. Mantovani, Detecting and locating non-technical losses in modern distribution networks. IEEE Trans. Smart Grid **9**(2), 1023–32 (2018)
19. P.O. Glauner, J.A. Meira, P. Valtchev, R. State, F. Bettinger, The challenge of nontechnical loss detection using artificial intelligence: a survey. Int. J. Comput. Intell. Syst. **10**, 760–775 (2017)
20. B. Yildiz, J. Bilbao, J. Dore, A. Sproul, Recent advances in the analysis of residential electricity consumption and applications of smart meter data. Appl. Energy **208**, 402–427 (2017)
21. S.K. Singh, R. Bose, A. Joshi, PCA based electricity theft detection in advanced metering infrastructure, in *7th International Conference on Power Systems (ICPS)*, Pune (2017), pp. 441–445
22. D. Dangar, S.K. Joshi, Normalization based K means clustering algorithm. IJREDR (2014)
23. E.W.S. dos Angelos, O.R. Saavedra, Detection and identification of abnormalities in customer consumptions in power distribution systems. IEEE Trans. Power Deliv. **26**(4) (2011)
24. J. Nagi, K. Yap, S. Tiong, S. Ahmed, M. Mohamed, Nontechnical loss detection for metered customers in power utility using support vector machines. IEEE Trans. Power Deliv. **25**(2), 1162–1171 (2010)
25. J. Nagi, K. Yap, S. Tiong, S. Ahmed, F. Nagi, Improving SVM-based nontechnical loss detection in power utility using the fuzzy inference system. IEEE Trans. Power Deliv. **26**(2), 1284–1285 (2011)
26. S. Aziz, S.Z. Hassan Naqvi, M.U. Khan, T. Aslam, Electricity theft detection using empirical mode decomposition and K-nearest neighbors, in *International Conference on Emerging Trends in Smart Technologies (ICETST)*, Karachi, Pakistan (2020), pp. 1–5
27. L. Connelly, Logistic regression. Medsurg Nurs. **29**(5), 353–354 (2020)
28. S. Chen, G.I. Webb, L. Liu, X. Ma, A novel selective naïve Bayes algorithm. Knowl. Based Syst. **192** (2020).
29. L. Li, S. Dai, Z. Cao et al., Using improved gradient-boosted decision tree algorithm based on Kalman filter (GBDT-KF) in time series prediction. J Supercomput **76**, 6887–6900 (2020)
30. M. Schonlau, R.Y. Zou, The random forest algorithm for statistical learning. Stata J. **20**(1), 3–29 (2020)
31. J. Pereira, F. Saraiva, A Comparative analysis of unbalanced data handling techniques for machine learning algorithms to electricity theft detection, in *IEEE Congress on Evolutionary Computation (CEC)*, Glasgow, United Kingdom (2020), pp. 1–8

32. Z. Qu, H. Li, Y. Wang, J. Zhang, A. Abu-Siada, Y. Yao, Detection of electricity theft behavior based on improved synthetic minority oversampling technique and random forest classifier. Energies **13**(8), 2039 (2020)
33. B.H. Shekar, G. Dagnew, Grid search-based hyperparameter tuning and classification of microarray cancer data (2019), pp. 1–8
34. P. Flach, Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. AAAI **33**(1), 9808–9814 (2019)