Tanzania Journal of Engineering and Technology 2022, **41(1)**:1-9
OPEN ACCESS articles distributed under Creative Commons Attribution Licence [CC BY-ND]
Websites: https://www.ajol.info/index.php/tjet/about; https://tjet.udsm.ac.tz

ISSN 1821-536X (**print**); ISSN 2619-8789 (**electronic**)
https://doi.org/10.52339/tjet.vi.767

Special Issue – First International Conference on Digital Transformation, 16 – 17, September 2021, Dar es Salaam, TANZANIA

Machine Learning Approach for Classifying Power Outage in Secondary Electric Distribution Network

Stephan Mgaya and Hellen Maziku

Department of Computer Science and Engineering, University of Dar es Salaam, P.O. Box 33335, Dar es Salaam, Email: smgayanath@gmail.com, nahelna@gmail.com

ABSTRACT

Power outage is the problem that hinders social and economic development especially for developing countries like Tanzania. Frequent power outages damage electric equipment, and negatively affect the industrial production process. Power outages cannot be completely eradicated due to uncontrolled cause like natural calamities but technical challenges can be managed and hence reducing power outages. The existing manual methods used to locate power outage like customer calls is inefficient and time consuming. On the other hand, modern method like the Advanced Metering Infrastructure (AMI) still faces a challenge in effectively classifying power line outage due to the nature of imbalanced datasets. Therefore, there is a need to develop a Machine Learning (ML) model to accurately classify power line outage. In this study, machine learning models are constructed from ensemble algorithms and tested using outage AMI data from 2012 to 2019 with 2 hours interval records. We propose the following ensemble-based machine learning approach to enhance classification; data sampling, algorithm weighting and finally ensembling. Results show that the Hybrid Stacking Ensemble Classifier (HSEC) model outperforms the others by accuracy of 0.981 G-mean, followed by Extra tree with accuracy of 0.964 G-mean. This model can be used in power line outage classification in any Secondary Electrical Distribution Network (SEDN). This study can be extended to locate power outage to household.

ARTICLE INFO

First presented: **16-17 September, 2021**

Submitted: **22 February, 2022**

Revised: **28 April, 2022**

Accepted: **May 20, 2022**

Published: **15 July, 2022**

Keywords: *Imbalanced dataset, Power outage location, Stacking ensemble classifier*

INTRODUCTION

The power outage is a crisis in any country whether it occurs on a small scale or large scale because it hinders social and economic activities in an area (Carreras *et al.*, 2003). Also, it affects several sectors and may cause severe damages in some entities and organizations such as hospitals, telecommunication facilities, and industrial installations.

Power outage is the loss of the electric power supply in part of the network or in the entire network thus affecting end-users. Power outage causes can be grouped into three main categories like hardware and technical failures, environmental incidents, and human errors (Sultan & Hilton, 2020). Electrical utility companies are required to use an Outage Management System (OMS) to identify the location where a power outage has occurred. This system allows the electric utility company to make restoration decisions like priorities based on the critical locations, size of the outage, and other factors. Having realized the significance of power supply, several measures have been implemented to cope with power outages, such as the installation of an OMS and smart meters to guarantee that power outages are detected and restored as soon as possible (Hoffmann & Payton, 2014).

The OMS is one of the most important systems for utility companies, when linked with Advanced Metering Infrastructure (AMI), geographic information system (GIS), Customer Information System (CIS), and Interactive Voice Response (IVR) can improve the functionality of OMS (Chakravarty *et al*., 2016). Countries have resorted to install smart meters in various regions to provide timely input in order to improve OMS.

The smart-meter, which plays a vital part in alleviating power outage problems using last gap massages, is a result of technological improvement in instrumentation. Addressing power outages with smart-meter data has become a widespread global practice because it minimizes identification of power outage duration by allowing for real-time and rapid delivery of power outage information, resulting in more effective power restoration efforts (Quilumba *et al*., 2015).

Moreover, the unplanned, scattered, and heterogeneous nature of the Secondary Electrical Distribution Network (SEDN) makes it challenging to locate power outages accurately at optimal time. Kumar & Pindoriya (2015) used AMI data to detect power outages in SEDN employing the fuzzy membership function probabilistic method of uncertainty algorithm to filter temporary and permanent outages. The authors also looked at a single service outage or multiple service outages using a fuzzy inference system model with the combination of AMI data and Supervisory Control and Data Acquisition (SCADA) information. An extension of the study suggests to facilitate location of power outages using AMI and Geographical Information System (GIS) data.

Kuroda *et al*. (2014) used distribution system model utilizing AMI data to determine the outage location in the distribution network using data collection status, live monitoring, and communication error monitoring. The results were visualized using the smart meters, transformers, and load bus icon approach. However, this approach takes time to collect faults data from time outage occurrence while not all power outages were caused by faults.

Xu & Chow (2006) studied the power outage causes using a fuzzy classification algorithm, and identified that the class imbalance issue encountered in many real-world scenarios including unbalanced power outages datasets as it often affects the performance of fault identification, especially for minority-class causes, since most commonly used methods aim to minimize the overall error rate. Classic Machine Learning (ML) algorithms such as support vector machines (SVM), artificial

neural networks (ANN), and decision trees (DT) usually have acceptable classification accuracy when using supervised technique assuming that the class distribution is balanced. But this assumption is often not realistic in real-world classification domains as occurrences of diverse classes shift significantly, which is known as the class imbalance problem (Wang *et al*., 2016). The power outage location dataset shows these qualities. Most existing ML algorithms fail to classify imbalanced class datasets due to bias in the majority class against the minority class. There are various proposed methods to address the class imbalance problem. These methods are based on data sampling techniques or modification of existing classifiers (Krawczyk, 2016).

Pristyanto *et al*. (2020) presented an ensemble model approach for dealing with class imbalance. The ensemble of algorithms included; Random Forests or Bagging Tree, Stacking Naïve Bayes and Decision Tree. The authors revealed that the ensemble model produced better performance than single classifiers. It is therefore concluded that the proposed algorithm has been unable to solve the outage location problem due to class imbalance. However, there are algorithms essential solving class imbalance problems that are necessary to be explored in addressing the situation.

This research presents the Hybrid Stacking Ensemble Classifier (HSEC) for dealing with imbalanced classes in classifying power line outage location. This method combines the best-performed base learner in stacking model under hybrid sampling to maximize the accuracy in classifying power line outage location using AMI dataset.

There are two contributions to our research. First, the HSEC model that we propose that can be used in power line outage classification in SDEN using AMI dataset. Second, the HSEC model that we propose can be a reference for further research related to handling class imbalances in the dataset in the field of machine learning.

METHODS AND MATERIALS

Addressing Imbalance

The methods addressing the class imbalance problems can be categorized into three groups: data-level methods, algorithmic-level methods, and ensemble methods.

Data level approaches

Data level approaches are based on resizing the training datasets to balance classes by using either over-sampling or under-sampling concepts. Under-sampling methods rebalance imbalanced class distribution by decreasing the number of majority class samples, whereas over-sampling methods reduce class disparity by generating new minority class samples. Dozens of over-sampling and under-sampling algorithms have been proposed. The common over-sampling algorithm is called SMOTE (Fernández *et al*., 2018). Among the under-sampling methods based on SMOTE are bSMOTE and V-synth. Modified under-sampling method is the Random Under-Sampling (RUS) that has been commonly used. There is also reported work using hybrid sampling that combines over-sampling and under-sampling approaches where majority classes are reduced and minority classes are increased to meet at the middle (Prachuabsupakij & Soonthornphisaj, 2014).

Algorithm level approach

Algorithmic level approaches focus on modifying existing classification algorithms to strengthen their ability to learn from minority classes based on SVM and Neural Networks. Ando (2012) proposed a new approach called nearest neighbor (NN) based on the k-Nearest neighbor density model to cope with the data imbalance issue. The main concern of NN is to employ an adjusted k-radius to compensate for the sparseness of the minority class. A wavelet support vector machine (WSVM) is presented. Concerning

imbalanced scenarios, a filter feature selection technique is performed to remove the redundant and irrelevant information. Datta & Das (2019) emphasized that the classical SVM moves the separating hyper-plane towards the minority class as the majority class is more likely to dominate the region of overlap. By this motivation, they proposed a Near-Bayesian Support Vector Machine (NBSVM) that utilizes Bayesian posterior probabilities to achieve the boundary shift as well as the unequal regularization costs.

Cost-sensitive algorithms attempt to increase the learning ability of classifiers by assigning larger misclassifying costs for minority class samples. López *et al*. (2012) proposed an algorithm to deal with large-scale imbalanced data using a fuzzy rule and cost-sensitive learning techniques. Krawczyk (2016) constructed a fusion algorithm based on cost-sensitive decision tree ensembles, in which the choice of cost matrix is estimated by Receiver operating characteristic (ROC) analysis. Thai-Nghe *et al*. (2010) introduced two empirical cost-sensitive algorithms, one combined sampling, cost-sensitive and SVM and the other treated the cost ratio as a hyper-parameter which needs to be optimized before training the final model. Another concept of learning from imbalanced classes is treating minority samples as outliers and analogizing technologies of detecting noises and outliers to model minority classes, such as one-class classifier.

Ensemble methods approach

Classifiers in ensemble tackle imbalanced learning improving significantly the performance of a single classifier. Ensemble methods can be viewed as building multiple classifier systems that combine a variety of base classifiers, for each base classifier data-level approaches which are often employed as a per-processing. The commonly used Multi Class System (MCS) is boosting algorithm proposed by Yijing *et al*. (2016) which has been applied in ensemble algorithms such as SMOTEBoost, RUSBoost, Easy Ensemble, EUSboost denoted that a typical MCS generally contains 3 processes, that are, re-sampling, ensemble building, and fusion rule. Sun *et al*. (2018) proposed ensemble strategy that converts an imbalanced dataset into multiple balanced subsets, where each subset a base classifier is trained. Krawczyk (2016) created an ensemble algorithm that contains sampling, pruning, and boosting technologies. The approach first divided the data into the non-overlapped region, borderline region, and overlapped region and then trained different regions by different classifiers. The imbalance situation of different amounts of data at the overlapped region and non-overlapped region is concerned. Zięba *et al*. (2014) proposed a boosted SVM, in which an active strategy of selecting the borderline examples to train each SVM is designed. In this way, each training set used to construct the basic classifier is more balanced and noiseless.

RESEARCH APPROACHES USED TO INVESTIGATE CLASSIFICATION

The study was conducted at Tanzania Energy Supply Company (TANESCO) in Dar es Salaam as it is the only electrical power utility company in the country. TANESCO manages generation, transmission, distribution, and operation of electrical systems network in Tanzania as the monopoly entity.

This study applied the mixed approach that combined qualitative and quantitative. The qualitative approach is used to interview TANESCO experts about the Distribution Management System (DMS), GIS, AMI, Service Delivery Management (SDM) to get a better understanding of how the electrical network works, causes of the power outage in the SEDN, methods used to identify and locate power outage, and restoration of power outages. The quantitative approach was used by taking

measurements collected from electrical instruments through the experiment to answer the research question on how power line outage location can be classified.

This study applied systematic algorithm development procedures to develop a Machine Learning algorithm to locate power outage line locations. The following are the systematic procedures used: Gathering algorithm requirements for estimating power outage location in SEDN, designing an algorithm based on the acquired requirements, Implementation of the algorithm, and evaluating the algorithm. These methods answered the research question purposely based on laboratory experiment setup and available data.

Qualitative approach

The qualitative data was collected through an interview approach. The interview was conducted in a focused group of TANESCO experts. This method is used to understand how power outage management work and how it is integrated with other systems like AMI, GIS, OMS, DMS to help locate power outages in SEDN.The interview revealed that there are two types of power outages, The planned power outages are caused by load shedding and unplanned power outage is caused by human activities, network problems, and others. Also, enlightened that customer call is the commonly used method in power outage location and finally the restoration process using restoration crew team.

Quantitative approach

The experimental data was a collected from Secondarily Distribution Transformer (SDT) AMI system. This data was collected for seven years from 2012 to 2019 with 20 minutes interval readings. This data was analyzed and then used in the proposed model to locate power line outages.

DEVELOPING ALGORITHM TO CLASSIFY POWER OUTAGE

The development of an algorithm is an experimental approach that can be used to identify power outage locations based on the AMI dataset. The AMI provides the voltage and current reading of each line from the transformer and can also send an outage notification when an outage occurs. The experimental set up used a Jupyter Notebook with python Scikit-learn libraries installed to develop the necessary algorithm following a number of stages.

The First stage involves preprocessing of the AMI dataset to handle missing values, labeling, and scaling of data before being trained. The second stage used seven Machine Learning to train prepared AMI datasets to have benchmark performance. These classical Machine Learning are SVM, Logistic Regression Classifier (LR), K-Neighbors classifier (KNN), Random Forest classifier (RF), Extra tree classifier (ET), AdaBoost classifier (AB), and Bagging classifier (BAG). The classical ML results were poor due to the imbalance of the dataset.

The third stage took four best-performing algorithms from the first stage to train the dataset under the data sampling technique. Moreover, the technique combined to produce a more robust algorithm classifier compared to individual techniques to deal with an imbalanced dataset.

Finally, the Hybrid Stacking Ensemble Classifier (HSEC) was proposed. A HSEC method is the combination of different base learners in the stacking model under a sampling technique. Some of these base learners are ensemble algorithms. This approach takes advantage of all the techniques discussed before to build stronger classifiers to deal with deployed imbalanced problems. The HSEC takes advantage of a strong base learner and aggregates using another meta learner to produced improved classification of skewed data. The dataset was re-sampled by using hybrid random sampling which is used to balance the imbalanced classes.

The HSEC uses the hybrid sampling technique that uses the over-sampling technique to generate more data in minority

classes and under-sampling to reduce data in the majority of the dataset to arrive at a balanced dataset at a common middle point. Figure 1 illustrates Hybrid Stacking Ensemble Model. The dataset is prepossessed then pass-through hybrid sampling to balance the imbalance dataset classes then subjected to HSEC. The Geometric Mean (*G-Mean*) is used as a metric in this study because it indicates the balance between classification performances on the majority and minority classes. Metric uses in machine learning depends on the nature of the problem. One of metrics used by other researchers is accuracy measure which is working for evaluating the performances of classifiers when dealing with balanced datasets in binary classification but is not appropriate for imbalanced datasets.

Lizárraga *et al*. (2008) recommended the need for metrics that consider each class's performance when dealing with the class imbalance; Where a poor performance of the positive instances gives a low G-mean value even if the negative instances are correctly classified by the model. The G-mean mathematical formula is defined in Equation (1).

$$G-Mean=\sqrt{recall\times specificity} \quad (1)$$

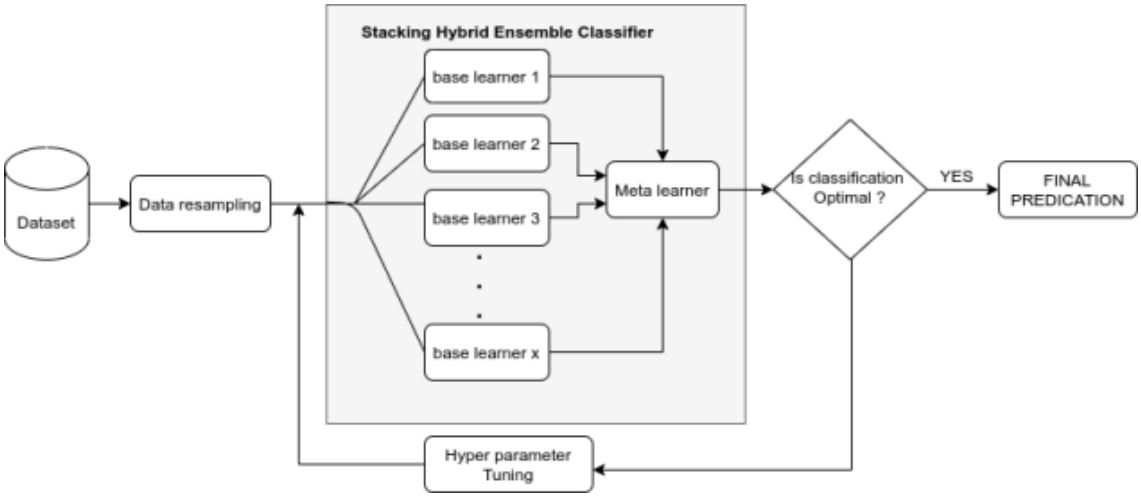

Figure 1: Hybrid stacked ensemble model.

RESULTS AND DISCUSSION

This section presents and discusses the results of different algorithms model performances on prediction of the outage location in the electrical distribution networks for imbalanced dataset. In this section different techniques such as re-sampling technique, cost estimation technique, and stacking ensemble are considered.

Results

In the first model, seven Machine Learning algorithms were used to classify the dataset. The objective was to compare performance of standard machine learning algorithms and provide a baseline classifier for further evaluations, i.e., applying data sampling to address the problem of imbalanced datasets. Figure 2 shows the result of the first experiment. The Random Forest classifier outperformed the rest by having an accuracy of 0.772 G-mean, therefore, the value becoming the benchmark for this study.

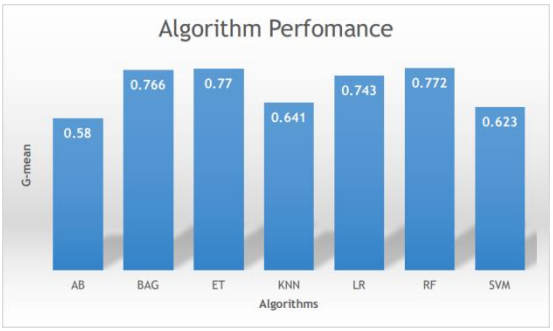

Figure 2: Illustrated classical ML results.

The AMI data was found to be highly imbalanced using skew and Kurtosis. Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails. Table 1 shows the skewness and kurtosis values. In this case, when the value of the skewness is negative, it means data is highly skewed. Kurtosis describes the peakiness of the distribution and is the measure of outliers present in the distribution. The dataset kurtosis value is 23, the high value of kurtosis in a dataset indicates that data has heavy outliers.

Table 1: Skew and Kurtosis results

	Skew	Kurtosis
L1	-4.92	23.51
L2	-4.92	23.50
L3	-4.90	23.35

The first quadrant in Figure 3 illustrates the class distribution of the original dataset meaning the frequency of power line outages is too small compared to the frequency when the power line is on.
Therefore, the data-sampling techniques were used to help deal with the dataset imbalance problem. Three types of sampling techniques were used over-sampling, under-sampling, and hybrid-sampling as visualized in Figure 3 in the second, third and fourth quadrant.

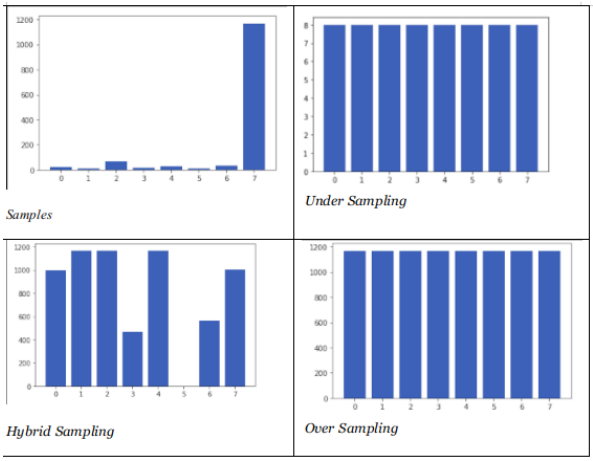

Figure 3: Data sampling result.

Table 2, presents the results of the second model that combines ML algorithms with sampling techniques to classify the power line outage location. Four well-performed algorithms were taken from the first experiment using the first model but configured with the K-Fold cross-validation (stratified) with k = 10 and repeat = 3. The algorithm was trained and evaluated 10 × 3 (30) times and each algorithm's mean and standard deviation were reported in Table 2. This configuration avoids fluke results and better captures the variance of the chosen model. Finally, the extra tree classifier under hybrid sampling had an overall better performance with the highest G-mean of 0.964.

Table 2: Model validation

Algorithm	*Sample*	*Under Sampling*	*Over sampling*	*Hybrid Sampling*
SVM	0.623 (0.061)	0.652 (0.115)	0.806 (0.047)	0.852 (0.022)
RF	**0.772 (0.060)**	0.726 (0.125)	0.923 (0.032)	0.941 (0.014)
BAG	0.766 (0.077)	0.730 (0.144)	0.904 (0.026)	0.931 (0.022)
ET	0.770 (0.052)	**0.795 (0.112)**	**0.939 (0.020)**	**0.964 (0.011)**

The final developed model used a stacking ensemble-based approach, this model combined the following algorithms (Support Vector Machine, Random Forest, Extra tree, and bagging classifier) as base learners and aggregate the result by using Logistic

regression under hybrid data sampling. The prediction was done under the same sampling technique and the results are presented in Figure 4. The results show that the stacking ensemble classification performed better than the other classifiers by 2%.

Hence, based on the results, stacking ensemble model outperformed other algorithms, having an accuracy of 98.1% G-mean. This concludes that a hybrid stacking ensemble is the best algorithm to classify power line outage location using any data sampling technique.

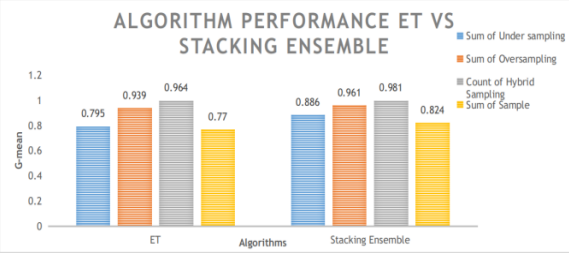

Figure 4: HSEC models performance under data sampling techniques.

CONCLUSION

The aims of this study were to classify the power line outage location in an electrical distribution network. The classification of power line outage location in the electrical power network based on imbalanced dataset HSEC has been done successfully. The results show that HSEC outperformed a single ML classier because of the contribution of the hybrid data sampling used to handle the imbalanced dataset classes.

RECOMMENDATIONS

This study was able to classify the power line outage location of SEDN using a HSEC. This study can be extended to identify the power outage to the individual household by integrating data from customer calls, AMI, and GIS.

REFERENCES

Ando, S. (2012). Performance-optimizing classification of time-series based on nearest neighbor density approximation. *Proceedings - 12th IEEE International Conference on Data Mining Workshops, ICDMW 2012*: 759–764. doi: 10.1109/ICDMW.2012.14.

Carreras, B.A., Lynch, V.E., Newman, D. E. & Dobson, I. (2003). Blackout mitigation assessment in power transmission systems, *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*: 10. doi: 10.1109/HICSS.2003.1173911.

Chakravarty, P., Rajsekar, V., & Ostrum, W. (2016). Integrating model-based OMS with ancillary systems. *2016 IEEE/PES Transmission and Distribution Conference and Exposition (T&D)*. doi: 10.1109/TDC.2016.7520020.

Datta, S., & Das, S. (2019). Multiobjective Support Vector Machines: Handling Class Imbalance with Pareto Optimality. *IEEE Transactions on Neural Networks and Learning Systems*, 30(5): 1602–1608. doi: 10.1109/TNNLS.2018.2869298.

Fernández, A., García, S., Herrera, F., & Chawla, N.V. (2018). SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. *Journal of Artificial Intelligence Research*, 61(2018): 863–905. doi: 10.1613/jair.1.11192.

Hoffmann, H., & Payton, D.W. (2014). Suppressing cascades in a self-organized-critical model with non-contiguous spread of failures. *Chaos, Solitons and Fractals*, *67*: 87–93. doi: 10.1016/j.chaos.2014.06.011.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4): 221–232. doi: 10.1007/s13748-016-0094-0.

Kumar, G., & Pindoriya, N.M. (2015). Outage management system for power distribution network. *2014 International Conference on Smart Electric Grid, ISEG 2014*. doi: 10.1109/ISEG.2014.7005598.

Kuroda, K., Yokoyama, R., Kobayashi, D., & Ichimura, T. (2014). An approach to outage location prediction utilizing smart metering data. *Proceedings - Asia Modelling Symposium 2014: 8th Asia International Conference on Mathematical*

Modelling and Computer Simulation, AMS 2014: 61–66. doi: 10.1109/AMS.2014.23.

Lizarraga-Lizarraga, G., Hernandez-Aguirre, A., & Botello-Rionda, S. (2008). G-Metric: An M-Ary quality indicator for the evaluation of non-dominated sets. *Association for Computing Machinery*: 665–672. doi: 10.1145/1389095.1389227.

López, V., Fernández, A., Moreno-Torres, J. G., & Herrera, F. (2012). Analysis of preprocessing vs. cost-sensitive learning for imbalanced classification. Open problems on intrinsic data characteristics. *Expert Systems with Applications*, 39(7): 6585–6608. doi: 10.1016/j.eswa.2011.12.043.

Prachuabsupakij, W., & Soonthornphisaj, N. (2014). Cluster-based sampling of multiclass imbalanced data. *Intelligent Data Analysis*, 18(6): 1109–1135. doi: 10.3233/IDA-140687.

Pristyanto, Y., Nugraha, A.F., Pratama, I., & Dahlan, A. (2020). Ensemble model approach for imbalanced class handling on dataset. *2020 3rd International Conference on Information and Communications Technology, (ICOIACT)*: 17–21. doi: 10.1109/ICOIACT50329.2020.9331984.

Quilumba, F.L., Lee, W.-J., Huang, H., Wang, D.Y., & Szabados, R.L. (2015). Using Smart Meter Data to Improve the Accuracy of Intraday Load Forecasting Considering Customer Behavior Similarities. *IEEE Transactions on Smart Grid*, 6(2): 911-918. doi: 10.1109/TSG.2014.2364233.

Sultan, V., & Hilton, B. (2020). A spatial analytics framework to investigate electric power-failure events and their causes. *ISPRS International Journal of Geo-Information*, 9(1): 54. doi: 10.3390/ijgi9010054

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with DTE-SBD: Decision tree ensemble based on SMOTE and bagging with differentiated sampling rates. *Information Sciences*, 425: 76–91. doi: 10.1016/j.ins.2017.10.017.

Thai-Nghe, N., Gantner, Z., & Schmidt-Thieme, L. (2010). Cost-sensitive learning methods for imbalanced data. *The 2010 International Joint Conference on Neural Networks (IJCNN)*: 1-8. doi: 10.1109/IJCNN.2010.5596486.

Wang, S., Minku, L. L., & Yao, X. (2016). Dealing with multiple classes in online class imbalance learning. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)*: 2118–2124.

Xu, L., & Chow, M.-Y. (2006). A classification approach for power distribution systems fault cause identification. *IEEE Transactions on Power Systems*, 21(1): 53–60. doi: 10.1109/TPWRS.2005.861981.

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, 94: 88–104. doi: 10.1016/j.knosys.2015.11.013.

Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing Journal*, 14(Part A): 99–108. doi: 10.1016/j.asoc.2013.07.016.