Electricity Theft Detection using Machine Learning

Ivan Petrlik[1], Pedro Lezama[2], Ciro Rodriguez[3], Ricardo Inquilla[4]
Julissa Elizabeth Reyna-González[5], Roberto Esparza[6]
Faculty of Industrial and Systems Engineering, National University Federico Villarreal, Lima, Perú[1, 2, 6]
Faculty of Electronic Engineering and Informatics, National University Federico Villarreal, Lima, Perú[3]
Faculty of Systems Engineering and Informatics, National University Mayor de San Marcos, Lima, Perú[3]
Faculty of Engineering, National University of Cañete, Cañete, Perú[4]
Faculty of Industrial and Systems Engineering, National University Hermilio Valdizán, Huánuco, Perú[5]

***Abstract*—This research work dealt with the indiscriminate theft of electric power, reported as a non-technical loss, affecting electric distribution companies and customers, triggering serious consequences including fires and blackouts. The research focused on recommending the best prediction model using Machine Learning in electrical energy theft. The source of the information on the electricity consumption of 42372 consumers was a dataset published in the State Grid Corporation of China. The method used was data imputation, data balancing (oversampling and under sampling), and feature extraction to improve energy theft detection. Five Machine Learning models were tested. As a result, the accuracy indicator of the SVM model was 81%, K-Nearest Neighbors 79%, Random Forest 80%, Logistic Regression 69%, and Naive Bayes 68%. It is concluded that the best performance, with an accuracy of 81%, is obtained by using the SVM model.**

***Keywords*—Energy theft; non-technical losses; machine learning; support vector machine**

I. Introduction

In the world, 70% of electricity consumption is lost and 30% in the Caribbean and South America, of which Peru stands out with 7% according to the Inter-American Development Bank [1]. Electricity losses are categorized into two categories: energy delivered to customers (unpaid energy) and losses generated in transmission and distribution lines, which are inherent to electricity transmission. Likewise, non-technical losses comprise the majority of losses in electricity networks and can account for more than 40% of the total electricity produced [2]. These types of losses are attributed to different sources, the most important and common being the alteration of metering equipment, illegal connections to the electrical grid, and energy theft [3]. Regarding distribution in Peru, annually, electricity theft generates losses of 103 million soles, equivalent to 207 GWh, for the companies providing the service [4]. However, this type of loss not only affects these companies but also the offenders themselves and people in the surrounding area, causing various accidents such as electric shocks, fires, and power outages.

According to [5], as presented in Table I, a division is made into countries, utilities, and society, which are categories represented in non-technical effects or consequences in which electric power has many losses.

The background of the respective research is based on multiple studies that have been conducted in many countries, designing intelligent systems that help to deal with this problem, mainly using Machine Learning techniques, which will be presented below:

In 2018, wide convolutional networks (CNNs) were used for one-dimensional data, and deep convolutional networks (CNNs) were used for two-dimensional data. The one-dimensional data were converted into two-dimensional electricity consumption data [6]. On the other hand, a study was carried out in which SVM was applied, using customer consumption data and the total energy distributed by the supplier, which allowed the calculation of the errors produced by electricity meters [7].

In 2019, a combination of neural networks, employed for the conversion of a one-dimensional dataset into a two-dimensional one, and random forests were used to perform customer classification [8]. In 2020, the k-nearest neighbors algorithm and empirical mode decomposition were used to extract the most important attributes from the dataset and obtain good accuracy in detecting energy theft [9]. Another study used a text convolutional neural network (Text-CNN) to effectively extract periodic features about energy consumption and detect electricity theft [10].

In 2021, several classification algorithms were compared, the main one being lightGBM, a fast algorithm based on decision trees, which achieved an accuracy of 84% [11]. Other algorithms compared are logistic regression, with an accuracy of 71%, stochastic gradient descent, with an accuracy of 65%, and decision tree, with 86%.

TABLE I. The Main Consequences of Non-technical Aspects of Power Theft

Countries	Utilities	Society
Increased use of scarce natural resources	Negative impact on the economy and finances	Total or partial outages
Increased contamination	Reduction of power plant efficiency	Increase in electricity rates
Increased use of public funds	Reduced capacity to upgrade the power system	Fires

According to [12], they proposed that for feature learning (classified into theft and non-theft), a deep convolutional neural network was used. Smart counters at different epochs provided data that was used for SVM training. The time interval of 15 minutes that the smart meter had to record the data through a source coming from the residential and industrial sector which is comprised of 26530 consumers which is the product of data collection.

In the study of [13], the authors evaluated 23 classifiers using the F1 score as a performance parameter. They used as a basis the data of a Brazilian company oriented to the electric power industry, with 261,489 consumers, with approximately 1400 fields. From the results obtained, they concluded that the classifiers (ensemble methods) are the most appropriate, allowing the identification of non-technical cases of electric power loss. The F1 score of 0.45 is the result of the gradient boosted three and an accuracy of 66.50% (actual field inspections) with respect to the rotation forest.

II. Materials and Methods

The procedure that was applied as a solution for electricity theft detection encompassed in the respective workflow is basically made up of five parts: data set acquisition, preprocessing, data balancing, feature extraction, classification, and acquired data set, shown in Fig. 1.

According to Fig. 1, the parts of the workflow will be detailed as follows:

A. Dataset Acquisition

The method of data collection was done through smart meters. The data comes from the daily consumption of electric energy belonging to the State Grid Corporation of China (http://www.sgcc.com.cn/), which was founded on December 29, 2002, and which supplies more than 1.1 billion inhabitants, covering 88% of the national territory. The description of the dataset used is presented in Table II.

According to Table II, we have the temporality range of the data, which comprises from January 1, 2014 to October 31, 2016 (approximately 147 weeks). The file size is 167 MB (175,194,613 bytes) in csv format , with respect to the data structure of the dataset is divided into customers who steal electricity amounting to 3615 (8.55%) and normal customers who consume electricity amounting to 38757 (91.5%) of which add up the total amount of records in 42372 (total customers).

B. Preprocessing

Usually, the electricity consumption represented by the dataset is constituted in some cases by erroneous and missing values and this is caused by problems in smart meters, storage with many problems, unreliability in transmission of metering data and others [20]. For the recovery of missing values in the content of the dataset of the respective research is the interpolation method [21] which is represented through the following formula 1:

$$f(x_i) = \begin{cases} \frac{(x_{i+1} + x_{i-1})}{2} & if\ x_i \in \text{NaN},\ x_{i-1}\ and\ x_{i+1} \notin \text{NaN} \\ 0 & if\ x_i \in \text{NaN},\ x_{i-1}\ or\ x_{i+1} \in \text{NaN} \\ x_i & if\ x_i \notin \text{NaN}. \end{cases} \quad (1)$$

Where:

x_i : Attribute of electricity consumption data

NaN: Non-numeric value

Next, the technique for the recovery of missing data was applied, using the average electricity consumption of each customer for which missing values were substituted. In addition, outliers were found, very different from the rest, which were restored using the equation of the three sigma rule, shown in the respective formula 2

$$f(x_i) = \begin{cases} avg(x) + 2 * std(x) & if\ x_i > avg(x) + 2 * std(x) \\ x_i & otherwise \end{cases} \quad (2)$$

where:

std(x) : typical deviation

avg(x): mean value of x

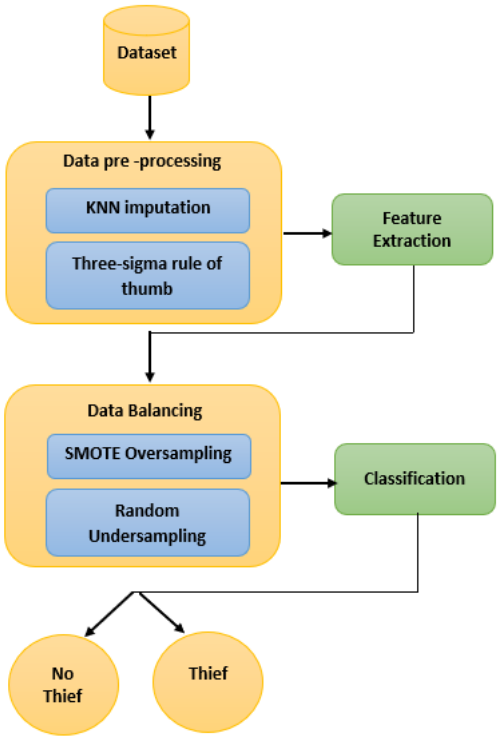

Fig. 1. Arch Workflow.

TABLE II. Description Del Dataset Utilizado

Description	Value
Temporal range of data	01/01/2014 – 31/10/2016
Dataset file size	167 MB (175,194,613 bytes)
Normal customers consuming electricity	38757 (91.5%)
Customers stealing electricity	3615 (8.55%)
Total customers	42372
Cases with missing data	Approximately 25%.

C. Feature Extraction

In order to classify the consumers, characteristics were extracted from their electricity consumption records. The characteristics used were the following: mean, standard deviation, peak to peak, skewness, median absolute deviation, entropy, and kurtosis.

D. Data Balancing

The dataset being used has been found to be imbalanced, with a greater amount of data representing people who are not stealing electricity compared to those who are stealing, which complicates the classification process. To balance the data, techniques such as oversampling and undersampling can be applied. For oversampling, a technique called SMOTE was often used, in which new instances are synthesized from other instances using the k-Nearest Neighbors technique [14]. However, it is suggested to use a subsampling technique in conjunction with the SMOTE technique [15]. For this research work on the dataset, the random subsampling technique was applied, dividing the data into disjoint training and test sets that are randomly partitioned several times [22].

E. Classification

According to [23], the classification process allows to obtain different classes, but based on a grouping of outputs through one or more input variables. In the research, a set of algorithms were applied for this purpose, each of which will be detailed below:

The SVM algorithm is designed to find the optimal separating hyperplane between classes based on support vectors (extremes of the class distributions). The training data are separated into classes using boundaries, which results in the maximization of the distance between the various data sets and the boundary [16].

The training dataset, consisting of n cases represented by {x_i, y_i}, i = 1...n, where y_(i)∈{1,-1}, is used to form a classifier for accurate generalization. A hyperplane is defined as:

$$w * x_i + b = 0 \quad (3)$$

Where there is a normal vector denoted by w and a point x, where both are in the hyperplane and b is the bias. And each point in the sample must satisfy:

$$y_i(w * x_i + b) > 1 \quad (4)$$

The k-Nearest Neighbors algorithm is a supervised classification algorithm that classifies or predicts based on proximity, which is calculated using various distance metrics [17]. In this study, we will use the Manhattan distance, defined as:

$$Manhattan\ Distace = d(x, y) = \sum_{i=1}^{m} |x_i - y_i| \quad (5)$$

Random forest is a supervised classification and regression algorithm that performs well on classification problems. It builds a set of decision trees and bases the final output on majority voting in classification problems. The decision tree algorithm for regression and classification is constructed by evaluating questions and node splits, which contribute to the further reduction of Gini impurities when answering [18].

Logistic regression is a classification algorithm that aims to predict or explain the values of a qualitative target variable as a function of a set of qualitative or quantitative explanatory variables. It is an extension of linear regression that uses the logit function for qualitative classification [19]. The logit function is defined as:

$$logistic(\eta) = \frac{1}{1+\exp(-\eta)} \quad (6)$$

Following the calculation of the conditional probabilities of which one event occurs with respect to the other, is the concept of Bayes Theorem of which naive bayes is a classification algorithm which is defined by the following respective formula:

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) * P(A)}{P(B)} \quad (7)$$

III. Results

The results were obtained based on the preprocessing of the data, totaling 33,009 instances, of which 20% were used for testing and 80% for training to predict the respective model to be compared. The processed dataset is shown in Table III.

TABLE III. Processed Dataset

Characteristic	Value
Total Instances	33009
Train Instances	26407
Test Instances	6602

A. Support Vector Machine Algorithm

After experimenting with different kernels to determine the optimal kernel for classification using the Support Vector Machine, the results of this experiment are shown in Table IV. The RBF (Radial basis function) kernel was chosen.

TABLE IV. Comparison between the Accuracies Obtained using Different Kernels

Kernel	Accuracy
Linear	75%
Polynomial	67%
RBF	80%
Sigmoid	72%

The parameters chosen were "gamma": 0.5 and "C": 100, obtaining an accuracy of 81%. Fig. 2 and Table V show the results through the classification report and the confusion matrix as follows.

B. K-Nearest Neighbors Algorithm

Using the Manhattan metric, the best number of neighbors for this classification was 5, as shown in Fig. 3, obtaining an accuracy of 79%. The results obtained in the confusion matrix and ranking report using these two parameters are defined in the following graphs (as seen in Fig. 4 and Table VI).

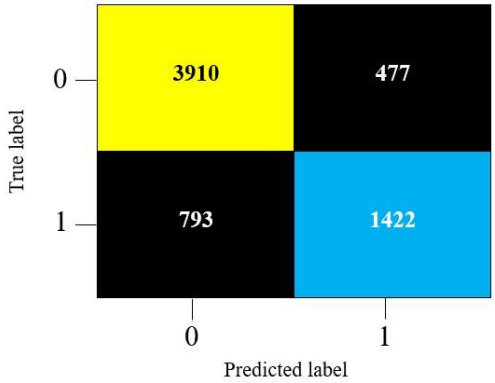

Fig. 2. Confusion matrix of the SVM based detection model.

TABLE V. SVM ALGORITHM CLASSIFICATION RESULT (DETECTION MODEL)

	Precision	Recall	F1 – score	Support
0	0.83	0.89	0.86	4387
1	0.75	0.64	0.69	2215
Accuracy			0.81	6602
Macro avg	0.79	0.77	0.78	6602
Weighted avg	0.80	0.81	0.80	6602

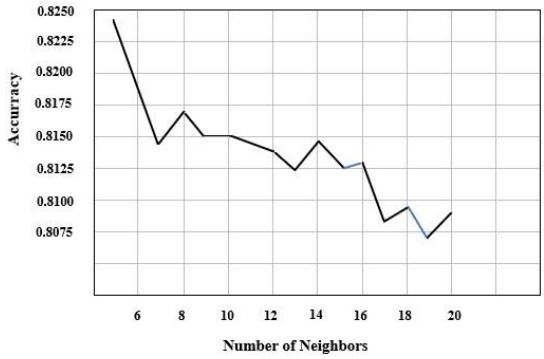

Fig. 3. Accuracies obtained using the manhattan metric and different numbers of neighbors.

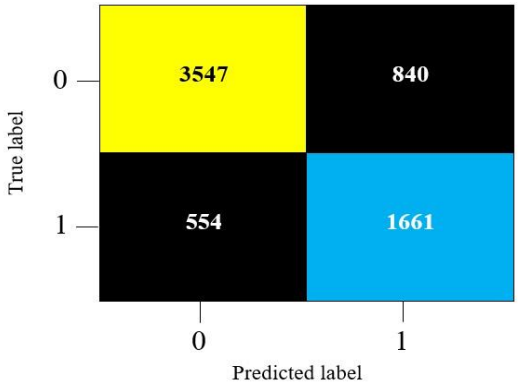

Fig. 4. Confusion matrix of the k-nearest neighbors based detection model.

TABLE VI. CLASSIFICATION RESULT OF THE K-NEAREST NEIGHBORS ALGORITHM (DETECTION MODEL)

	Precision	Recall	F1 – score	Support
0	0.86	0.81	0.84	4387
1	0.66	0.75	0.70	2215
Accuracy			0.79	6602
Macro avg	0.76	0.78	0.77	6602
Weighted avg	0.80	0.79	0.79	6602

C. Random Forest Algorithm

The parameters chosen were 'max_depth': 20, 'n_estimators': 100, 'max_figures': 'auto', 'criterion'='entropy', achieving an accuracy of 80%. The following classification report and confusion matrix are determined by the following graphs (Fig. 5 and Table VII).

The following classification report and confusion matrix were generated using the parameters 'max_depth': 20, 'n_estimators': 100, 'max_figures': 'auto', 'criterion': 'entropy', which resulted in an accuracy of 80%. The results are shown in Fig. 5 and Table VII.

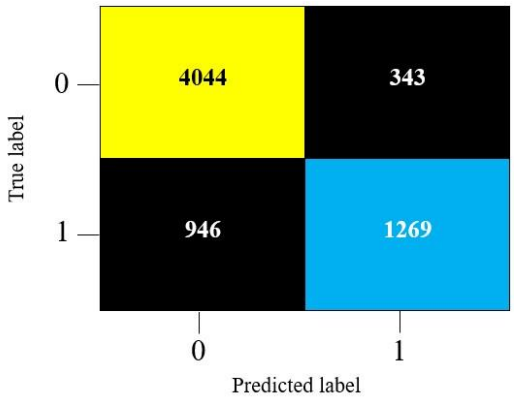

Fig. 5. Confusion matrix of the random forest based detection model.

TABLE VII. RANDOM FOREST ALGORITHM CLASSIFICATION RESULT (DETECTION MODEL)

	Precision	Recall	F1 – score	Support
0	0.81	0.92	0.86	4387
1	0.79	0.57	0.66	2215
Accuracy			0.80	6602
Macro avg	0.80	0.75	0.76	6602
Weighted avg	0.80	0.80	0.80	6602

D. Logistic Regression Algorithm

This model was trained using 1000 iterations and the inverse of the regularization strength 'C' as 10, obtaining an accuracy of 69%. The classification report and the confusion matrix results were obtained, which are determined through the following graphs (Fig. 6 and Table VIII):

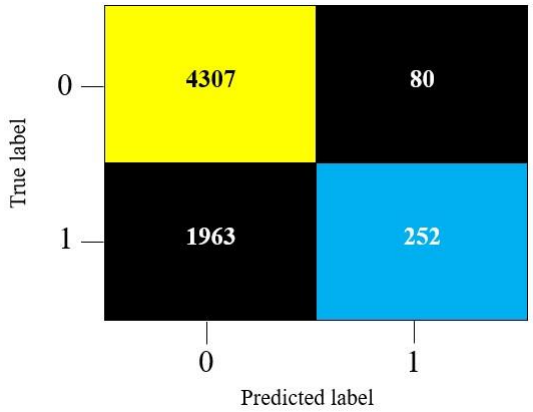

Fig. 6. Confusion matrix of the logistic regression based detection model.

TABLE VIII. LOGISTIC REGRESSION ALGORITHM CLASSIFICATION RESULT (DETECTION MODEL)

	Precision	Recall	F1 – score	Support
0	0.69	0.98	0.81	4387
1	0.76	0.11	0.20	2215
Accuracy			0.69	6602
Macro avg	0.72	0.55	0.50	6602
Weighted avg	0.71	0.69	0.60	6602

E. Naive Bayes Algorithm

The default parameters for classification were used for this algorithm, obtaining the following results (see Fig. 7 and Table IX).

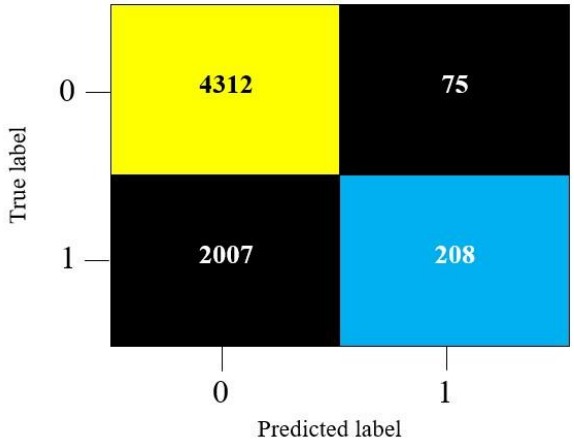

Fig. 7. Confusion matrix of the naive bayes based detection model.

TABLE IX. CLASSIFICATION REPORT OF THE NAIVE BAYES BASED DETECTION MODEL

	Precision	Recall	F1 – score	Support
0	0.68	0.98	0.81	4387
1	0.73	0.09	0.17	2215
Accuracy			0.68	6602
Macro avg	0.71	0.54	0.49	6602
Weighted avg	0.70	0.68	0.59	6602

Table X shows a consolidation of the results obtained from the percentage values of the accuracy indicator of all the proposed models.

TABLE X. LIST OF RESULTS OF THE ACCURACY VALUES OF THE PROPOSED MODELS

Machine Learning Models	Accuracy
Support Vector Machine	81%
Random Forest	80%
K-Nearest Neighbors	79%
Logistic Regression	69%
Naive Bayes	68%

As shown in Table X, the SVM model has a higher accuracy indicator score of 81%.

IV. DISCUSSION

The research of [6] focused on making a comparison of CNN, SVM, LR, RUSBoost models in order to know who has the best prediction. The accuracy result of the SVM model was 0.772, contrasting with our research that also developed a comparison of models such as SVM, RF, KNN, LR and NB, having the best accuracy results of 0.81 for SVM and 0.80 for RF. If we compare the SVM model results of both researches, there is an improvement of 0.038 (3.8%) in favor of the present research. Likewise, the research of [8], also makes a comparison of models such as CNN-RF, CNN-GBDT, CNN-SVM, CNN, SVM, RF, LR and GBDT, the SVM model has an accuracy of 0.77, compared with the present research, achieving an improvement of 0.04 (4%). Next, we have another research by [10], which proposes a new model (TextCNN) for electricity theft detection and also makes a comparison with traditional machine learning models (LR, SVM), the SVM model has an accuracy of 0.70, compared with the present research, achieving an improvement of 0.11 (11%). The SVM model has been compared for all research, however, this model compared with the research [6], which uses the CNN model, results in an accuracy of 0.92 (92%) and the research [10], whose model is Text-CNN whose accuracy value is 0.90 (90%), although it is true that both have better performance, however more computing power is needed when identifying consumers who steal electricity.

V. CONCLUSIONS

This research proposed an electricity theft detection model based on Support Vector Machine using electricity consumption information obtained from the State Grid Corporation of China, achieving a maximum detection accuracy of 81%.

The models have limitations because it was not possible to correctly classify about 25% of the electricity theft cases, which may be due to the lack of data on electricity thieves compared to those who did not steal electricity. However, we attempted to solve this problem using data balancing techniques (oversampling and under sampling).

The experiments conducted show that the system using SVM performs better than most of the other prediction

systems tested, such as Logistic Regression, Random Forest, and K-Nearest Neighbors, while the Naive Bayes model does not correctly fit this problem.

REFERENCES

[1] R. Jiménez, T. Serebrisky, and J. Mercado, Power Lost: Sizing Electricity Losses in Transmission and Distribution Systems in Latin America and the Caribbean. Inter-American Development Bank, 2014. doi: 10.18235/0001046.

[2] P. Glauner, P. Valtchev, C. Glaeser, N. Dahringer, R. State, and D. Duarte, "Non-Technical Losses in the 21st Century: Causes, Economic Effects, Detection and Perspectives," 2018. [Online]. Available: https://www.researchgate.net/publication/325297875.

[3] B. K. Hammerschmitt et al., "Non-Technical Losses Review and Possible Methodology Solutions," Proceedings - 2020 6th International Conference on Electric Power and Energy Conversion Systems, EPECS 2020, pp. 64–68, Oct. 2020, doi: 10.1109/EPECS48981.2020.9304525.

[4] "Hurto de Energía - enel.pe." https://www.enel.pe/es/ayuda/hurto-de-energia.html (accessed Nov. 17, 2021).

[5] de S. Savian, J. C. M. Siluk, T. B. Garlet, F. M. do Nascimento, J. R. Pinheiro, and Z. Vale, "Non-technical losses: A systematic contemporary article review," Renewable and Sustainable Energy Reviews, vol. 147, p. 111205, Sep. 2021, doi: 10.1016/J.RSER.2021.111205.

[6] 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, Apr. 2018, doi: 10.1109/TII.2017.2785963.

[7] S. C. Yip, W. N. Tan, C. K. Tan, M. T. Gan, and K. S. Wong, "An anomaly detection framework for identifying energy theft and defective meters in smart grids," International Journal of Electrical Power & Energy Systems, vol. 101, pp. 189–203, Oct. 2018, doi: 10.1016/J.IJEPES.2018.03.025.

[8] S. Li, Y. Han, X. Yao, S. Yingchen, J. Wang, and Q. Zhao, "Electricity Theft Detection in Power Grids with Deep Learning and Random Forests," Journal of Electrical and Computer Engineering, vol. 2019, 2019, doi: 10.1155/2019/4136874.

[9] S. Aziz, T. Aslam, S. Zohaib, H. Naqvi, and M. U. Khan, "Electricity Theft Detection using Empirical Mode Decomposition and K-Nearest Neighbors," 2020.

[10] X. Feng et al., "A novel electricity theft detection scheme based on text convolutional neural networks," Energies (Basel), vol. 13, no. 21, Nov. 2020, doi: 10.3390/en13215758.

[11] S. V. Oprea and A. Bâra, "Machine learning classification algorithms and anomaly detection in conventional meters and Tunisian electricity consumption large datasets," Computers and Electrical Engineering, vol. 94, Sep. 2021, doi: 10.1016/j.compeleceng.2021.107329.

[12] U. Haq, J. Huang, H. Xu, K. Li, and F. Ahmad, "A hybrid approach based on deep learning and support vector machine for the detection of electricity theft in power grids," Energy Reports, vol. 7, pp. 349–356, Nov. 2021, doi: 10.1016/J.EGYR.2021.08.038.

[13] R. M. R. Barros, E. G. da Costa, and J. F. Araujo, "Evaluation of classifiers for non-technical loss identification in electric power systems," International Journal of Electrical Power and Energy Systems, vol. 132, Nov. 2021, doi: 10.1016/J.IJEPES.2021.107173.

[14] He and Y. Ma, "Imbalanced learning: Foundations, algorithms, and applications," Imbalanced Learning: Foundations, Algorithms, and Applications, pp. 1–210, Jan. 2013, doi: 10.1002/9781118646106.

[15] N. v. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal Of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2011, doi: 10.1613/jair.953.

[16] Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Predicting Student Success Using Big Data and Machine Learning Algorithms," International Journal of Emerging Technologies in Learning (iJET), vol. 17, no. 12, pp. 236–251, Jun. 2022, doi: 10.3991/IJET.V17I12.30259.

[17] "What is the k-nearest neighbors algorithm? | IBM." https://www.ibm.com/topics/knn (accessed Apr. 26, 2022).

[18] R. Sujatha, S. L. Aarthy, J. M. Chatterjee, A. Alaboudi, and N. Z. Jhanjhi, "A Machine Learning Way to Classify Autism Spectrum Disorder," International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 6, pp. 182–200, 2021, doi: 10.3991/IJET.V16I06.19559.

[19] T. Hamim, F. Benabbou, and N. Sael, "Survey of Machine Learning Techniques for Student Profile Modelling," International Journal of Emerging Technologies in Learning (iJET), vol. 16, no. 4, pp. 136–151, 2021, doi: 10.3991/IJET.V16I04.18643.

[20] C. Genes, I. Esnaola, S. M. Perlaza, L. F. Ochoa, and D. Coca, "Recovering missing data via matrix completion in electricity distribution systems," in Signal Processing Advances in Wireless Communications (SPAWC), 2016 IEEE 17th International Workshop on, 2016.

[21] Zheng, Z.; Yang, Y.; Niu, X.; Dai, H.N.; Zhou, Y. Wide and deep convolutional neural networks for electricity-theft detection to secure smart grids. IEEE Trans. Ind. Informat. 2017, 14, 1606–1615. [CrossRef]

[22] L. Rokach, Ensemble learning: Pattern classification using ensemble methods (second edition). World Scientific Publishing Company, p.2010, 2019.

[23] R. D. K. Hiran, Machine learning: Master supervised and unsupervised learning algorithms with real examples. New Delhi, India: BPB Publications, 2021.