energies

MDPI

Article

Identification of Nontechnical Losses in Distribution Systems Adding Exogenous Data and Artificial Intelligence

Marcelo Bruno Capeletti [1], Bruno Knevitz Hammerschmitt [1], Renato Grethe Negri [2], Fernando Guilherme Kaehler Guarda [3], Lucio Rene Prade [4], Nelson Knak Neto [5,*] and Alzenira da Rosa Abaide [1]

[1] Graduate Program in Electrical Engineering, Federal University of Santa Maria, Santa Maria 97105-900, Rio Grande do Sul, Brazil
[2] Technologic Center, Federal University of Santa Maria, Santa Maria 97105-900, Rio Grande do Sul, Brazil
[3] Santa Maria Technical and Industrial School, Federal University of Santa Maria, Santa Maria 97105-900, Rio Grande do Sul, Brazil
[4] Polytechnic School, University of Vale dos Sinos, São Leopoldo 93022-750, Rio Grande do Sul, Brazil
[5] Academic Coordination, Federal University of Santa Maria, Cachoeira do Sul 96503-205, Rio Grande do Sul, Brazil
* Correspondence: nelson.knak@ufsm.br

Abstract: Nontechnical losses (NTL) are irregularities in the consumption of electricity and mainly caused by theft and fraud. NTLs can be characterized as outliers in historical data series. The use of computational tools to identify outliers is the subject of research around the world, and in this context, artificial neural networks (ANN) are applicable. ANNs are machine learning models that learn through experience, and their performance is associated with the quality of the training data together with the optimization of the model's architecture and hyperparameters. This article proposes a complete solution (end-to-end) using the ANN multilayer perceptron (MLP) model with supervised classification learning. For this, data mining concepts are applied to exogenous data, specifically the ambient temperature, and endogenous data from energy companies. The association of these data results in the improvement of the model's input data that impact the identification of consumer units with NTLs. The test results show the importance of combining exogenous and endogenous data, which obtained a 0.0213 improvement in ROC-AUC and a 6.26% recall (1).

Keywords: nontechnical losses; outliers identification; exogenous data; power system distribution; artificial neural networks; hyperparameter optimization; data mining; big data

check for updates

Citation: Capeletti, M.B.; Hammerschmitt, B.K.; Negri, R.G.; Guarda, F.G.K.; Prade, L.R.; Knak Neto, N.; Abaide, A.d.R. Identification of Nontechnical Losses in Distribution Systems Adding Exogenous Data and Artificial Intelligence. *Energies* **2022**, *15*, 8794. https://doi.org/10.3390/en15238794

Academic Editor: Andrzej Bielecki

Received: 18 October 2022
Accepted: 16 November 2022
Published: 22 November 2022

1. Introduction

Nontechnical losses (NTL) are characterized by theft, fraud or irregularities in the consumption of electricity that are not properly or effectively billed by distribution companies. In addition to financial losses attributed to the costs of stolen energy, NTLs have consequences for the quality of energy delivered to consumers that can, in the worst case, result in material damage.

NTLs can be carried out in several ways, sometimes ingenious. The most common and easily practiced energy theft occurs when the consumer performs a by-pass of the meter clock, thus forging or not accurately counting the consumer units (CUs). However, other methods are used by fraudsters. In [1], the authors describe that meter violations are a major cause of NTLs in addition to other methods used by fraudsters, such as tampering with voltage or current measurement or applying a magnetic field or high-frequency waves.

In this context, NTLs are a common problem in all electrical systems around the world, and they mainly focus on electric distribution systems, where voltage levels are adjusted for the end use of electricity. In addition, NTLs are directly related to the socioeconomic indices of a nation, and consequently, an even greater problem in developing regions such as Latin America and sub-Saharan Africa [2]. Therefore, the mitigation of NTLs warrants constant study, so that the identification of fraud, theft or irregularity results in financial returns on the part of energy companies and reductions in consumers' costs as well as maintaining the network standards within the established parameters.

Within this scenario, data science and artificial intelligence (AI) are areas of great expansion and development in the last decade, and their main objectives are to detect patterns and generate knowledge through data [3]. When used assertively, they can be applied to solve complex and nonlinear problems such as identifying outliers in historical data series, as in the case of CU with NTL.

AI models aim to develop algorithms to perform cognitive tasks, with attributes capable of storing knowledge through experience and employing them to design solutions [4]. One of these models is artificial neural networks (ANN), regarded as one of the most modern and powerful machine learning (ML) and AI tools. The ANN is inspired by the human brain and capable of operating massively parallel and distributed systems, as well as performing nonlinear and highly complex mappings, with the possibility of adjustment through hyperparameters.

Today, ANN with ML and deep learning (DL) are state-of-the-art models of NTL detection. As an example, convolutional neural networks [5], long short-term memory (LSTM) [6], and hybrid models [7,8] together with other binary classification models are regularly applied in research and have achieved the best robustness in the classification of electricity consumers with NTL.

One important characteristic of a multilayer perceptron (MLP) ANN is the ability to learn complex and nonlinear relationships between data and targets, which allows the employment of categorical and continuous variables in the same level. In addition, there is no restriction on data format. In [7], a MLP ANN was used in a hybrid model where it received several inputs such as longitude and latitude and voltage from the CU. It was also used as part of the hybrid model in [9], where geographic information data were passed to the MLP. Next, in the search for architectural optimization, it was used in the last hidden layers of [10] to detect NTLs. However, the treatment and quality of the database is of vital importance for the good performance of NTL classification models.

However, the problem of identifying NTLs is characterized by great imbalance between classes, since the anomalies (fraudulent consumer) occur at a much lower frequency than the normal data (regular consumer or N-NTL). Therefore, it is important to emphasize that the algorithm requires amounts of data representative of the two classes [11]. There are techniques and methods for mitigating the effects that occur in the model due to the unbalanced database (DB). The techniques are mainly used in the learning phase of the classification algorithm, as it becomes necessary to have representation of both classes of consumers, as in the case of data oversampling and undersampling.

These sampling methods are reviewed in [12], which compared several oversampling and undersampling techniques considering recall as a relevant performance metric. The authors concluded that undersampling majority data is the best sampling technique. Likewise, the strategy used to label the data of each CU in [13], which compares between labeling data by analyzing the expertise in the area and by means of field inspections, through the results of technical inspections of suspected consumers. The authors [13] concluded that the second approach had the best results. Additionally, ref. [14] proposed a method for generating suspicious consumers knowing the patterns of fraudster consumers and regular consumers in a model trained using simulated samples.

Recently, new concepts are being developed in the field of ML, among them the application of exogenous data, that is, information external to the environment under analysis, to improve the performance of models. Additionally, concepts of essential size (smart size) and data-centric models have also been highlighted [15]. In essence, the concepts aim to improve the quality and quantity of data, so that they contribute to the robustness of the ML project.

In [16], a classification proposal was developed to identify suspicious energy consumption profiles, comparing these with regular energy consumption profiles. For this, the distance between the customer with suspicious consumption and regular consumption is calculated. In [17], comments from inspectors and energy company staff, consumption information and contract information such as contracted power and economic activity were used as input resources. Similarly, in [18], socioeconomic status as a categorical variable was used for each electrified sector, and activity was classified as high, medium or low. In [19], econometric models are estimated to analyze and predict short-term NTLs. The dataset contains number of inhabitants in risk areas, default rate, and the time variable, representing the year of data collection. Table 1 summarizes and highlights some references related to the identification of NTL that employs ANN models.

Table 1. Summary of references related to the identification of NTLs.

Reference	Inputs	ANN Model	Description
[6]	Mainly consumption (sequential)	LSTM	The model detects abnormalities in a consumer's power consumption and classifies it as an anomaly, uses advanced metering infrastructure and consumption with half-hour intervals.
[7]	UC longitude and latitude and voltage	Hybrid (MLP-LSTM)	Uses a LSTM model for the consumption of smart meters and auxiliary data are the input for the MLP; the model outperforms other models in PR-AUC and ROC-AUC.
[8]	Consumption (daily smart data)	Hybrid (CNN-LSTM)	Uses consumption data to characterize consumers as either benign or thieves and a synthetic minority oversampling.
[9]	Geographic information	Hybrid (MLP-GRU)	The author highlights fraudster users' generation techniques for the problem of imbalance and the use of auxiliary data.
[10]	Consumption (Daily)	Optimized	Uses Bayesian optimization to find the optimal neural network structure.
[20]	Multiple Sources	Multiples	Employs various data sources such as climate and location to identify sources of electricity.
[21]	Consumption	Hybrid (LSTM–UNet–Adaboost)	Uses SMOTE with an imbalanced dataset obtained from a smart grid.

A study carried by [22] show that exogenous data can be useful as additional information for improve load forecasting accuracy. This study evidences that several factors can influences on power consumption, especially weather data, but also others can be related to consumption of electricity like socio-economic and demographic information. As can be seen, the ANN is recognized and widely used in classification models due to its easy comprehension, configuration, and reproduction, especially with qualitative and quantitative data [23]. More detailed theoretical aspects of the ANN-MLP that is used in this research can be found in [24].

It is well known that temperature significantly influences energy consumption. Many studies have approached the problem, such as [25,26]. In [27], there is a strong correlation between consumption and temperature, especially in the residential market. In addition, in [20], it was observed that illegal consumers use more electricity than legal consumers and that the consumption of electricity by these consumers intensifies in periods of high and low temperatures, in order to establish thermal comfort conditions. In [28], an applied tool was developed for the selection of climatic variables for load forecasting. In this study, heat caused people to turn up their air conditioners, and cold caused them to turn up their heaters. In [29], an index was defined for possible climate effects on electrical energy irregularities: As this index increases, diverging from the considered ideal point that is assumed for a comfortable temperature, electrical energy consumption should increase. Table 2 summarizes and highlights some references related to the validation and use of temperature in NTL identification.

Table 2. Temperature Validation.

Author	Description
[20]	The consumption of electric energy is intensified in times of high and low temperatures.
[28]	Due to thermal discomfort from cold and heat, there is a tendency to increase demand on of the system.
[29]	In electrical energy irregularities, energy consumption increases, diverging from the point considered ideal that is assumed for a comfortable temperature.
[30]	Focuses on fraudulent activities based on the temporal analysis of consumers who show reduced consumption in summer and winter.

Due to the various fraud modes reported in the electricity sector, managing NTLs is one of the most challenging tasks for energy distribution companies. Therefore, studies that use AI and ML models are essential for the identification of outliers so that energy companies can dispatch technicians to locations more likely to find NTLs, since depending on the method used by the fraudster, the determination and confirmation of the fraud can be costly and technically difficult.

Thus, for the development of tools for detecting NTL, several sources of exogenous data can be useful as input objects of the system, each one of these sources of data being related to climatic factors such as temperature. Even though there is scant evidence in the literature of the use of exogeneous data [20] such as temperature in a model of identification of NTL, the unavailability of weather data for every place or granularity in the database (hourly/daily/monthly) could lead to a misclassification of NTL consumers, by applying mismatching data. To overcome that, a previous study [31] presents the use of the correlation of ambient temperature data associated with electricity consumption, where consumption is the main data point for the identification of NTL [3]. The results show that the use of this metric can bring significant improvement in NTL detection models at the same time that it can reduce the probability of the misclassification of NTL consumers by using temperature only as a support variable.

Therefore, this work proposes the use of exogenous data in a model of the identification of NTL. The exogenous database used in this study is of climate data; specifically, the monthly average temperature is the support input variable of the problem. There is also a real database of an electricity distribution company in southern Brazil, Electrical Energy State Company—Companhia Estadual de Energia Elétrica (CEEE) and Equatorial Group—Grupo Equatorial. In order to use the exogenous variable, Pearson correlation and Euclidean distance were calculated, these variables being inputs to the classification model in addition to the endogenous data from the energy company. Data processing was performed, such as undersampling and temporal **matching** of **endogenous and exogeneous** data. The classification model used was the MLP, in which hyperparameters were optimized to improve the architecture's performance. The results were compared with the model without exogenous data, where a significant improvement in the robustness of the model with exogenous data could be observed.

The main contributions of this research are:

- The development of an innovative framework for the identification of NTL on distribution networks based on the introduction of exogeneous data (temperature);

- The introduction of the temporal matching of endogenous and exogeneous databases in order to establish correlation between consumption behavior and temperature variation based on the same set point (last inspection);
- The introduction of Pearson correlation and Euclidian distance as inputs of the ANN MPL, applying temperature as a support variable, then reducing the probability of the misclassification of NTL consumers;
- The analysis of the impact of using temperature to identify NTL consumers.

The paper is organized as follows: Section 2 presents the metrics and adjustments for the development of this paper and the definitions for implementing the model. In Section 3, the results and discussion are shown, and the classification model developed. Finally, in Section 4, the paper concludes.

2. Methods

This work proposed a NTL detection framework based on endogenous and exogenous data to identify CUs with NTLs. For this, a supervised classification approach was used, which characterizes it as a data-oriented method, developed from the optimized MLP model.

The methodology described in this work was developed in Python programming language using the ANACONDA package in the Spyder Integrated Development Environment. The libraries used were: Pandas, NumPy, MatPlotLib, Sk-Learn, Imbalanced-Learn, Keras, TensorFlow and Keras-Tuner. The developed model was implemented on a workstation with Intel(R) Xeon(R) Silvar 4108 CPU @1.8 GHz (2 Processors), 64 GB of RAM, and a Microsoft Windows Server 2016 operating system.

The general flowchart of this work can be seen in Figure 1, which explains in a simplified way the model developed, which can be subdivided into 4 modules called: extraction, mining, AI, and evaluation.

The first module has the function of extracting raw data from the concessionaire's database, as well as from the National Institute of Meteorology (INMET) [32]. It also performs temporal standardization for monthly basis of temperature data. Thus, the temperature data, which are available at hourly intervals, are standardized on a monthly basis, through the calculation of the monthly average, adapting them to the same temporal granularity as the consumption data.

The second module is essential for properly modeling the problem. Initially, the CUs were labeled as NTL (1) or regular (0). Then, manipulating the results of inspections previously carried out by the distribution company, the unit was labeled, and the date of the inspection was logged. This was necessary to match the starting time of the time series and the exogeneous data. After that, temporal data mining was performed with the objective of temporally adjusting the CU data from the previously saved inspection date. In addition, missing and/or null data (DB error) in the consumption database were identified, and a new record was attributed through the average of the existing data in the time series. As a result, all consumption and temperature records were adjusted for the 36 months prior to the inspection date. Subsequently, Pearson's correlation, Euclidean distance, the standard deviation, the probability distribution function (PDF), the median, and the amount of zero consumption were obtained. Finally, the training, validation, and test databases were separated.

Then, in the third module, the training and validation of the ML model was carried out together with the optimization of the architecture and hyperparameters. For the optimization of the MLP, the values within a search range were optimized to identify the best architectures. Regarding the model's hyperparameters, the number of hidden layers, number of neurons per layer, and dropout rate were optimized.

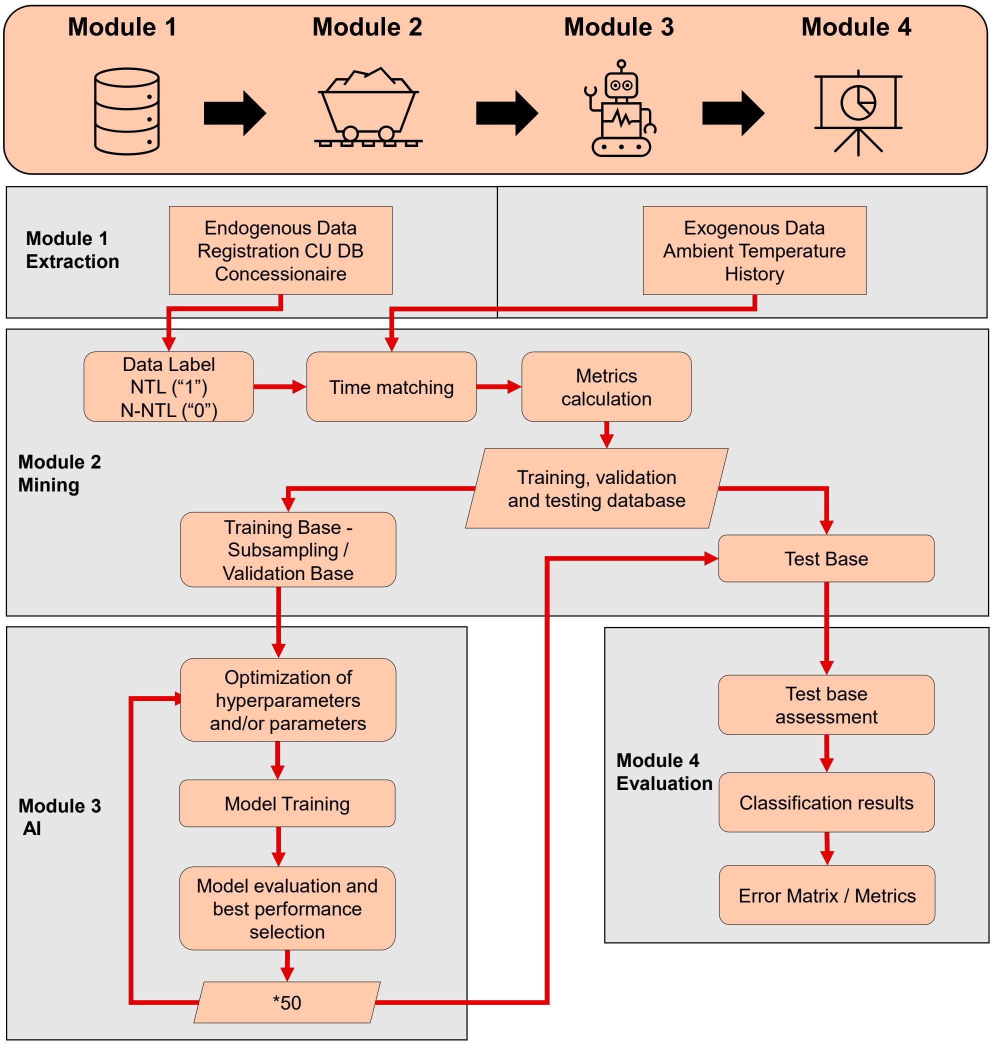

Figure 1. Flowchart of the proposed methodology. *50 means that Hyperparameter optimization 50 training iterations.

Finally, the fourth module evaluated the results obtained by the models of the previous stage, through the error matrix, accuracy calculation, error rate (recall) and precision of the results of the model’s test base, in addition to the graphical analysis of the ROC curve (receiver operating characteristic), from the PR (precision–recall) curve, and from the AUC (area under the curve) of the ROC and PR curves. The processes of the developed model are described in detail below. As the extraction module has already been fully described, the extraction details will not be covered.

2.1. Data Mining

The data mining stage for this study comprised labeling the CUs, treating the metrics that would serve as input for the proposed classification model, making adjustments, and dividing the database into training, validation, and testing sets. The procedures for mining the DB are explained in the next sections.

2.1.1. Inputs Adjustments for Neural Networks Models

In general, inputs must be preprocessed for correct entry into the intelligent system, as ML algorithms learn through experience. For this, it is essential that the input data used for learning be as faithful as possible, expressing in fact the consumption patterns of consumers with NTL and regulars.

This work proposes a temporal adequacy for consumers inspected in the field. As the main objective was to identify different consumption patterns and the changes in consumers' behaviors influenced by an inspection, adjusting the time series was necessary to meet the same starting time. The last time the consumer was inspected by the distribution company was taken as the reference. It means that the date on which the inspection was carried out is identified in order to match the starting time of the consumption and temperature time series. From that, the time series size is set to 3 years (36 months) prior to the inspection date, as shown in Figure 2.

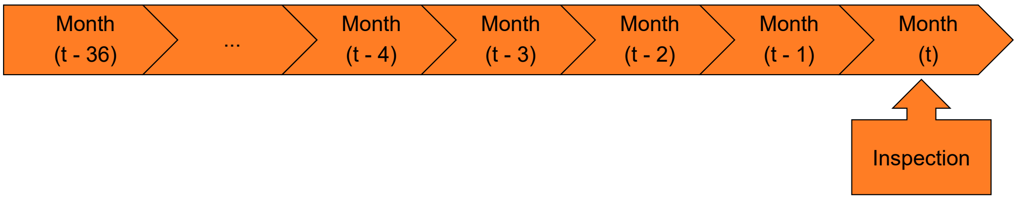

Figure 2. Mining temporal series: Consumption and Temperature.

This temporal matching of consumption and temperature inputs, taking as reference the last inspection, had the objective of properly defining the behavior of a CU with NTLs in the period before the inspection is carried out, that is, the behavior of the consumer during the period in which the irregularity was committed. This manipulation is highlighted because as exemplified in Figure 3, the behavior of a CU with NTLs can change soon after the irregularity is identified and rectified. Therefore, the algorithm must be trained with data that clearly show the behavior of the CUs during the period of irregularity in order to identify the NTL consumption patterns. It is noteworthy that the units were labeled through the results of field inspections previously carried out by the energy company.

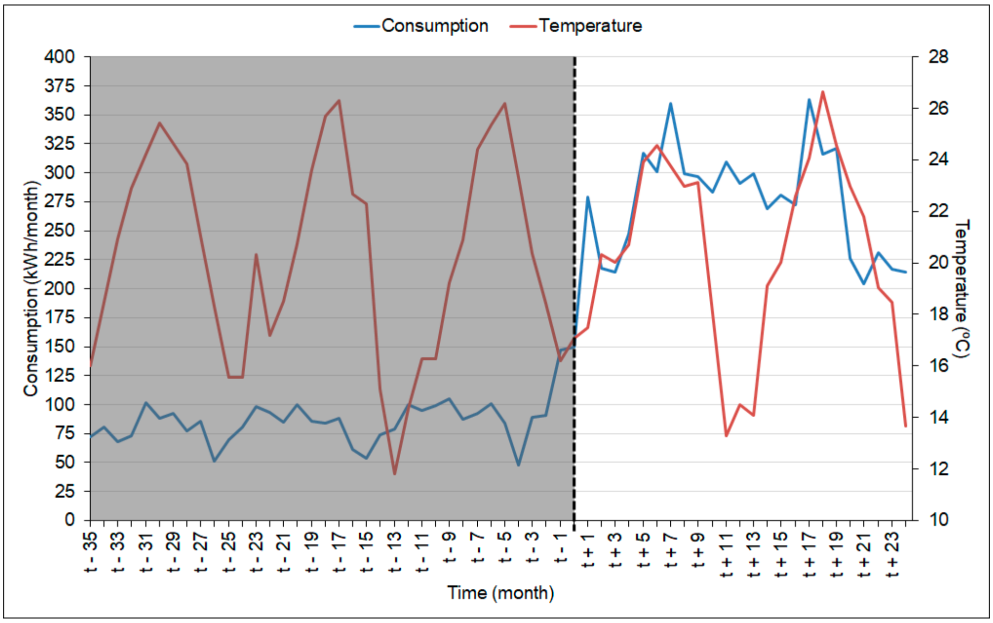

Figure 3. Consumer with NTL (fraudulent behavior highlighted).

2.1.2. Summary, Similarities, and Distribution Metrics

To analyze the consumption of CUs, a series of summary metrics are used to synthesize the information contained in a time series and thereby summarize the information contained on the DB. The first one is the arithmetic mean ($\overline{x}$), exemplified in Equation (1):

$$\overline{x} = \frac{1}{n} \sum\nolimits_{i=1}^{n} x_i, \tag{1}$$

where x_i are the monthly consumption samples of the CU and n the total of samples contained in the series.

The next metric is the median $\widetilde{x}$, which depicts the central value of the CU consumption samples (x_n). For the calculation of this metric, if the total of the samples is even, $\widetilde{x}$ is calculated by Equation (4), as the average of the two central values x_{n1} e x_{n2}; otherwise it is calculated by the second function:

$$\widetilde{x} = \frac{x_{n1} + x_{n2}}{2} \ or \ \widetilde{x} = x_n, \tag{2}$$

Then there is the standard deviation (s) of the CU consumption samples, denoted by Equation (3). This is the most important measure of dispersion as it provides a measure related to the mean. It is also important to note that the standard deviation is strongly affected by the presence of anomalies:

$$s(x) = \sqrt{\frac{1}{n-1} \sum\nolimits_{i=1}^{n} (x_i - \overline{x})^2}, \tag{3}$$

To use the PDF, an adaptation was made in the way in which the indexes obtained were treated. For this, the maximum (C_{max}) and minimum consumption (C_{min}) were identified within the consumption history of each CU under analysis. Subsequently, it was necessary to sort the data history into intervals of the same amplitude, followed by $C_0 - C_1$, $C_1 - C_2$, … , $C_{n-1} - C_n$, with C_0 equal to C_{min}, and C_n equal to C_{max}. After delimiting the intervals, the number of samples that fell within each interval were counted, thereby identifying the interval with the largest number of samples. For this interval, the average of the consumption samples that fit there was calculated by Equation (1), being called $\overline{x}_{PDF}$. From this, it was possible to more faithfully portray the consumption profile of this CU, making it its identity.

To assess the similarity of temperature and consumption, this work employs two similarity metrics: Euclidean distance and Pearson correlation. To determine the collinearity between monthly temperature and monthly consumption of the CUs, it is assumed that the greater the collinearity of these two variables, the lower the probability of NTLs in the CU under analysis. In fact, the introduction of these metrics makes the temperature itself work as a support. Both Pearson correlation and Euclidian distance are the temperature-related inputs for the ANN MPL. That assures the establishment of features correlated to consumption and temperature, then reduces the probability of any misclassification of NTL consumers.

The Euclidean distance represents the physical distance between points in a space; it also consists of the shortest distance between the points in the dataset. Therefore, it is manifested by a continuous line between the samples of the consumption dataset, that is, it can demonstrate similarity between two independent vectors. This metric is suitable for applications that do not necessarily present a correlation between different measures, as soon as it compares variables of the same temporal location, expressed by Equation (4):

$$d(x,y) = \sqrt{\sum\nolimits_{i,j=1}^{n} (x_i - y_j)^2}, \tag{4}$$

where d is the Euclidean distance, x_i is the first vector under analysis (consumption) and y_j the second vector (monthly average temperature data).

It can be noted that the Euclidean distance is invariant when dealing with changes in the order of time; this means that it does not capture vector correlation, for which other similarity measures are used such as Pearson's correlation [33]. Pearson's correlation (r) corresponds to an absolute value situated between +1 and -1 that reflects the intensity of the relationship between two data vectors, which can be classified as positive (with an interval of $0 < r \leq 1$) or negative (with an interval of $-1 \geq r > 0$). A positive correlation expresses that the data are directly proportional, and a negative correlation also expresses that the variables are correlated but inversely proportional. If the Pearson correlation is 0, there is no relationship between the datasets [34,35]. Pearson's correlation is described by Equation (5), which results in the degree of relationship between consumption and temperature:

$$r(x,y) = \frac{\sum_{i,j=1}^{n}(x_i - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \cdot \sum_{j=1}^{n}(y_j - \overline{y})^2}}, \tag{5}$$

where r is the absolute Pearson correlation, x_i the CU consumption data, $\overline{x}$ the arithmetic mean of the CU consumption samples, y_j the monthly average temperature data, and $\overline{y}$ the historical average temperature.

Finally, the normal distribution of each metric listed as input data of the model is made, demonstrated by Equation (8), which was used to standardize the inputs, scaling to a unitary variation:

$$z(x_i) = \frac{x_i - \overline{x}}{s(x_i)}, \tag{6}$$

where z is the normal distributed value of the standardized metric, x_i is the individual reference sample, $\overline{x}$ the arithmetic mean of this same metric, and s the standard deviation of the values of vector x_i.

2.1.3. Data Adjustments

The number of CUs with NTLs and their proportion to regular consumers are important values for detecting outliers in a supervised approach. In order to improve the robustness of the classification, it is necessary for the model to be able to capture patterns from CUs with NTLs and regular consumers. Additionally, the temporal adequacy increases the number of CUs labeled as irregular as the time interval corresponding to field inspections can be defined; that is, old inspections can be used with minimal addition of noise in the outlier detection model. This allows the model to learn from the consumption profiles of CUs with older irregularities and even increase the database with more CUs labeled as NTLs. It is noteworthy that these adjustments were used in the training, validation, and testing databases of the model.

As a result of the temporal adequacy of the time series described, and with the use of the results of field inspections dated from January 2020 to May 2022, CUs with NTLs varied from January 2017 to April 2022, depending on the month in which the inspection was carried out. For consumption time series, it is still necessary to treat null data and missing records. For this, within the necessary period of 3 years, if the record of a given month did not exist or was null, the record was inserted as the average of the existing records.

Then, the parameters related to the consumption database were calculated, with the median (2), standard deviation (3), PDF, and the amount of zero consumption. Parameters involving the exogenous temperature variable and consumption data were also obtained with Euclidean distance (4) and Pearson correlation (5). It is noteworthy that these parameters were obtained from data previously adjusted in time. Similarly, inspections that took place between January 2020 to May 2022 were used for the CUs.

After the aforementioned procedures, and according to the mining module in Figure 1, the database was divided between training, validation, and test sets, defined respectively with the percentages of 80%, 10%, and 10%, as established in [7]. Then, the data were undersampled in the training base, in which the undersampling equaled the number of

CUs labeled as NTLs or regular, taking into account that NTL consumers are found in much lower numbers than regular ones, as indicated in Figure 4.

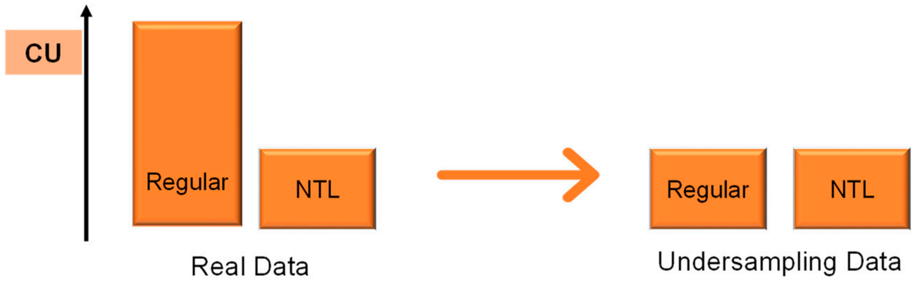

Figure 4. Undersampling—Training database.

In addition, it was necessary to make adjustments to the categorical and numerical variables in the entire database. Table 3 presents numerically the size of each of the entries in the ANN model. Each of the categorical inputs went through a coding process, where the categorical variables were replaced by numeric variables with values between zero and the number of classes of the variables minus one, using the label encoder algorithm.

Table 3. Model inputs size.

Input	Size
Consumption	36
Correlation and Euclidian Distance	2
Standard deviation, PDF and median	3
Quantity of zero consumption	1
City	1
Neighborhood	1
Main class description	1
Connection Type	1
Perimeter	1
Contracted Voltage Level	1
Street type	1

2.2. Multilayer Perceptron Neural Networks

To implement the ML model, the MLP algorithm was used through the TensorFlow library [36]. It is important to mention that due to its extension, the theoretical aspects of MLP are not presented. If necessary, more detailed information about MLP can be found in [24].

All categorical and time series variables need to go through a normalization process according to Equation (6) and are later used as inputs in the ANN according to Figure 5. It is noteworthy that the time-series normalization process was carried out aiming to remove imperfections from the time series and attenuate the differences in scales that were obtained from the monthly consumption measurements.

Then it was necessary to define the activation functions (f) of the hidden and output layers. In this work, the linear function (*Relu*) was used in the hidden layers, and the sigmoid function was used in the output layer, where the functions are shown in Figure 6, this choice being defined with the objective of reducing the error of the MLP. Furthermore, the sigmoid function has limits between 0 and 1, so the output of the ANN varies between these limits, being thus described as the probability that the CUs will be evaluated as consumers with NTLs.

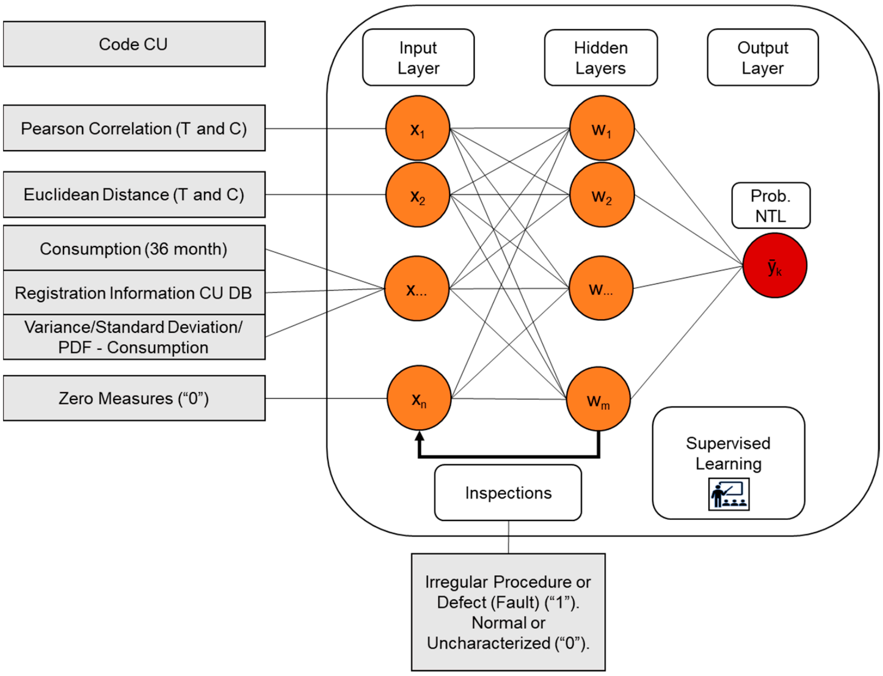
Code CU
Pearson Correlation (T and C)
Euclidean Distance (T and C)
Consumption (36 month)
Registration Information CU DB
Variance/Standard Deviation/ PDF - Consumption
Zero Measures ("0")
Input Layer
Hidden Layers
Output Layer
x1
x2
x...
xn
w1
w2
w...
wm
Prob. NTL
ŷk
Supervised Learning
Inspections
Irregular Procedure or Defect (Fault) ("1"). Normal or Uncharacterized ("0").

Figure 5. Inputs and MLP.

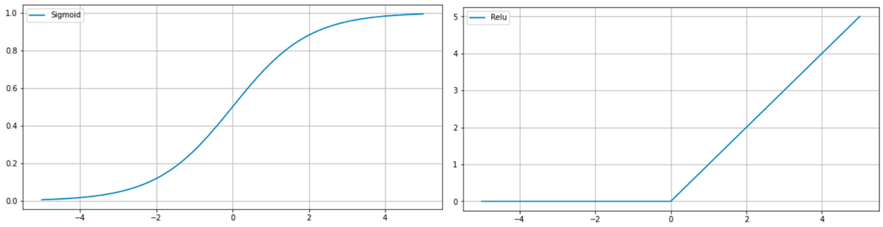

Figure 6. Linear and Sigmoid Functions.

The MLP training error calculation optimization method used in this work was proposed by [37] and widely used in the literature for optimization in binary classification ANN. The algorithm losses were calculated through the binary entropy function of Equation (7), which calculates the cross-entropy loss between true labels and predicted labels. This method of calculating the cross-entropy loss is mostly used in binary classification algorithms (0 or 1):

$$L_p = -\frac{1}{N}\sum_{i=1}^{N} y_i * \log(pred(y_i)) + (1 - y_i) * (\log(1 - pred(y_i)) \tag{7}$$

Optimization

Hyperparameter optimization is one of the most important steps in the MLP learning process, in which it was possible to optimize the number of hidden layers along with the number of neurons in each layer of the neural network and also the dropout rates of neurons from each MLP layer. The batch size was defined as the number of samples (CUs) that propagate through the MLP before the weights are updated. Finally, dropout rates were defined as the function to eliminate neurons from the layers randomly, aiming to reduce the severity of overfitting.

For the optimization of the model, 50 combinations of hyperparameters of the MLP were trained, observing the parameters of Table 4 and aiming at the ROC-AUC metric, which was optimized during 50 training iterations. The model was trained with a batch size of 1, configuring an online training and with a number of 50 Epochs. The ANN training ended when the maximum number of iterations of the model was reached, aiming at the lowest loss value (7) of the validation base. Then, the model with the best performance was used for the evaluation module, where the classification of the data from the previously separated test base was performed.

Table 4. Hyperparameter optimization–Search space.

Parameters	Search Space
Number of hidden layers	[1–5]
Number of neurons	[12–64]
Dropout	[0.1–0.5]

2.3. Model Evaluation

To evaluate the robustness of the classification model, one can have 4 evaluations of the output, portrayed by the error matrix or confusion matrix, that evaluate the outputs of the model, namely: false positives (*FP*), originally false with positive outputs; true positives (*TP*), originally positive with positive outputs; false negatives (*FN*), originally negatives with negative outputs; and true negatives (*TN*), originally negatives with positive outputs, as in [11]. It is noteworthy that the evaluations of *TP* and *TN* should be maximized and *FP* and *FN* minimized. Next, in the error matrix of Table 5, 1 is portrayed as CUs with NTLs and 0 as regular CUs:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

However, for an unbalanced DB, the accuracy metric should not be used alone, as one can find predominant negative true values given that for an unbalanced DB, the negative values of the bank are the vast majority. Therefore, accuracy alone does not clearly demonstrate how the classifier is behaving for the anomalous class (CUs with NTL). In view of this, the robustness of this data class has little impact on the metric [38]. Therefore, other metrics are needed that, associated with accuracy, will support a more reliable evaluation of the model, as observed in [8,21].

Table 5. Error Matrix.

		Predicted Class	
		0	**1**
True Class	**0**	*TN*	*FP*
	1	*FN*	*TP*

Thus, other evaluation parameters for an unbalanced DB and quantified by the error matrix are the hit rate over the real data obtained by Equation (9), also known as recall, and the accuracy of the classification model in the prediction of 0 and 1, expressed by Equation (10). In addition, the ROC curve, the ROC-AUC, PR curve, and the PR-AUC are used to analyze the results of classification, and it is necessary to compute the metrics, in addition to the *Recall* (9) and *Precision* (10), *False Positive Rates* (11), and *True Positive Rates* (12).

$$Recall = \frac{TN}{TN + FP} \ or \ \frac{TP}{TP + FN} \tag{9}$$

$$Precision = \frac{TN}{TN + FN} \ or \ \frac{TP}{TP + FP} \tag{10}$$

$$False\ Positive\ Rates = \frac{FP}{FP + TN} \tag{11}$$

$$True\ Positive\ Rates = \frac{TP}{TP + FN} \tag{12}$$

Additionally, the ROC is a curve obtained through the threshold variation of evaluation of the ML model results, changing the false positive and true positive rates. Thus, a classifier that produces continuous outputs such as the MLP can vary according to the output limit between its extremes from 0 to 1, resulting in the ROC curve as shown in [39], which illustrates the classifier's inability to avoid false positives and ability to detect the positive class correctly [40].

Therefore, from the ROC curve, the ROC-AUC metric can be obtained, which is one of the most appropriate metrics for evaluating ML results in unbalanced DBs and detecting outliers. This metric ranges from 0.5 to 1, where 0.5 is a bad classification value and 1 is an ideal classifier value. Several studies use this metric for evaluation, as observed in [7,38,41], and based on this, it was used to evaluate the performance of the developed model.

The PR curve was also used, which evaluates the results of precision and recall, respectively, in Equations (11) and (12). Therefore, by the graphical analysis of the PR curve, it was possible to evaluate the performance of the model, observing the elevation of the curve, because the performance increased together with the PR curve [7]. Therefore, as the ROC curve was obtained from the ROC-AUC, the PR-AUC was obtained from the PR curve, which aims to evaluate the average precision of the rates of predicted positive values and the rate of true positives. Thus, with the graphic analysis and the ROC-AUC and PR-AUC metrics, greater reliability is given to the model proposed in this study.

It is important to notice that classification models based on ML techniques are built from a specific database. Therefore, the origin, quality, and size of the database are influenced by the problem being analyzed. The same happens to the time-series granularity or the sampling strategy adopted. It makes the comparison among classification models a difficult task because depending on the availability of data, different strategies can be considered to evaluate the performance of the classifier over unseen data [42]. Therefore, these metrics are commonly used as reference for the evaluation of the model [7,20].

In addition, the main focus of this research was not the model itself but the innovative framework that is based on the employment of exogeneous and endogenous data, being either categorical or continuous, in order to identify NTL consumers. The MLP is recognized and widely used in classification models due to its easy comprehension, configuration and reproduction, especially in cases of the employment of qualitative and quantitative data [7].

3. Results and Discussion

For the validation and testing of the methodology, it was determined to use a database provided by CEEE and Equatorial Group, which exclusively counts previously inspected CUs. In this way, greater reliability was obtained in the results since all CUs were labeled according to the real field results, where it was possible to accurately label whether or not there was an incidence of NTL.

Then, the steps of the methodology up to the mining module of Figure 1 were applied, with the adjustments previously defined, and null and nonexistent consumption time-series data were replaced. Subsequently, the remaining missing data from the DB were filtered, resulting in 111,500 CUs that were then used in the training, validation, and testing phases of the model.

3.1. Database Characterization

In order to describe the CUs of the DB used in this study, it is appropriate to analyze the number of unique values described in all CUs. With this, it is highlighted that CUs from 13 cities were used, totaling 2275 neighborhoods, which are classified by the type of connection (single-phase, two-phase, or three-phase connection) and by the voltage level the energy is provided (four different groups). Table 6 shows the overall information of database.

Table 6. General Inspected Consumer Units.

Inputs	Unique Values
City	13
Neighborhood	2275
Connection Type	3
Contracted Voltage Level	4

It is important to notice that all cities, neighborhoods, connection types, and contracted voltage levels must be introduced to the model as samples of regular or NTL CUs. As presented, the ML model obtains knowledge through experience, where it is necessary to have representation of the most varied NTL patterns to map such patterns and accurately identify irregularities. To obtain that, the database was split in two groups labeled "Regular" for CUs that did not present NTLs or "NTL" if NTLs had been previously identified. Then, both groups were sized considering unique values for each input. The dimensions of the labels of the CUs of the DB used in this study are shown in Tables 7 and 8. This information was essential to further sampling.

Table 7. Inspected Consumer Units Labeled Regular.

Inputs	Unique Values
Cities	13
Neighborhoods	1675
Connection Type	3
Contracted Voltage Level	4

Table 8. Inspected Consumer Units Labeled as NTL.

Inputs	Unique Values
Cities	13
Neighborhoods	2076
Connection Type	3
Contracted Voltage Level	4

For the analysis of the complete CU DB, CUs were grouped by the average through the previously defined 0 or 1 labels. In this way, mean (1), median (2), standard deviation (3), PDF, Euclidean distance (4), and Pearson's Correlation (5) were grouped, thus making it possible to analyze the distinction between regular CUs and NTLs. As can be seen in Table 9, the CUs labeled as NTLs have lower average consumption, a higher standard deviation, a lower median of consumption, and a lower PDF compared with CUs labeled as regular.

Table 9. Description of CUs Consumption Labeled as NTL and Regular.

Variables	Regular (0)	NTL (1)
Average consumption (kWh)	197.14	159.44
Consumption standard deviation	67.65	69.60
Consumption median	187.55	149.78
Consumption PDF	178.12	139.47
Euclidian distance	1175.95	986.04
Pearson correlation	0.1606	0.0787

The graphical analysis of the frequency distribution of the normalized units in relation to Pearson's correlation is expressed in Figure 7. The observed trend depicts the clustering of more regular CUs with higher correlations when compared with irregular CUs. In an analogous way, analyzing the histogram of Figure 7, the frequency distribution of irregular units is concentrated in greater proportion for negative correlations and differs from regular units. Therefore, it is possible to affirm that the behavior of regular CUs differs from the behavior of irregular CUs, this being explicit when we include the temperature as a comparison bias and energy consumption through Pearson's correlation.

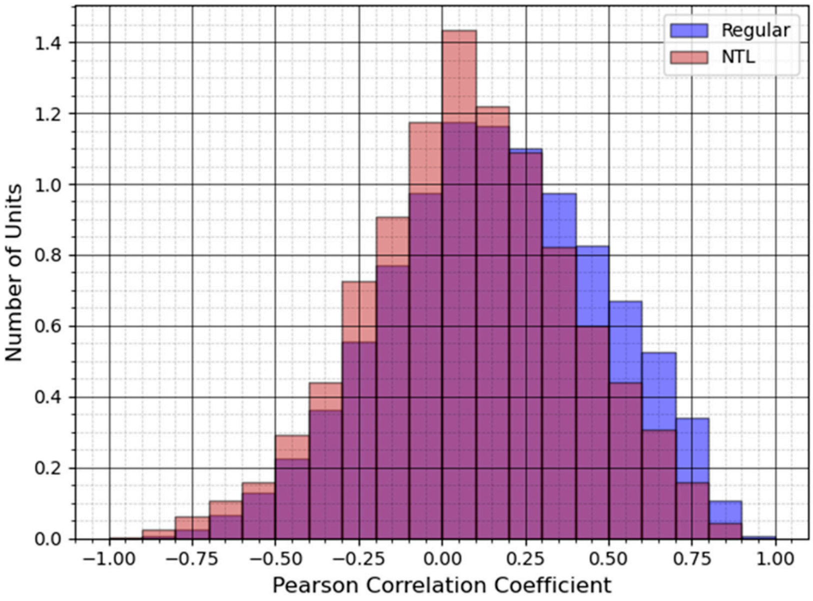

Figure 7. DB Pearson correlation distribution.

Now, the histogram has been produced, and it is expected to obtain improvement in the robustness of the NTL identification model, in which the temperature is inserted as an input variable, through the calculation of Pearson's correlation between consumption and temperature. The same can be seen in Figure 8, which depicts the frequency distribution of the normalized units in relation to the Euclidean distance. Although not very expressive, more information is acquired for the smallest normalized values of this variable, which are added to the model classification results.

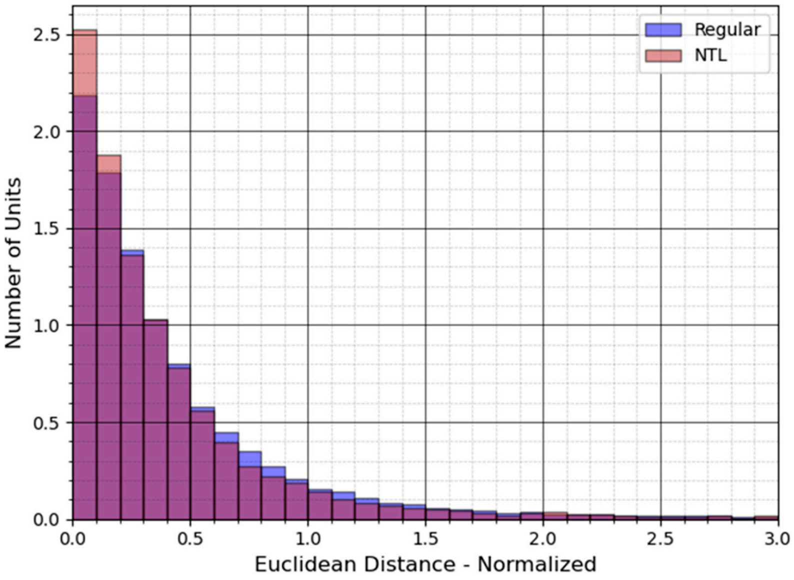

Figure 8. DB Euclidian distance distribution.

3.2. ANN Optimized Model

For use in the classification model, the complete DB had 84,682 regular CUs and 26,818 CUs with NTLs. That is, 24.05% of samples had NTLs. After labeling the CUs, the database was separated into training, testing, and validation datasets, which were taken from the complete database at random. The training base has 80% of the complete base, made up of 67,746 regular CUs and 21,454 CUs with NTL. As previously explained, this base went through a balancing process that resulted in a training base with 21,454 CUs and 21,454 CUs with NTLs. As for the validation and test sets, both comprised 10% of the complete DB, which represented 8468 regular CUs and 2682 CUs with NTLs in each set, as shown in Table 10.

Table 10. Training, validation and testing DB.

Base	Regular CU	NTL CU	%NTL
Train	21.454	21.454	50%
Validation	8.468	2.682	24.05%
Test	8.468	2.682	24.05%

3.3. Results Analysis

For the evaluation of the optimized ANN, the test dataset was used. It is noteworthy that the test set has a total of 11,150 CUs, with 8468 CUs labeled as regular and 2682 CUs labeled as irregular CUs or with NTLs.

The output neuron of the ANN has the sigmoid activation function, and the output of this neuron fluctuates between 0 and 1, thus making it possible to generate thresholds for the evaluation of the results. This threshold can be adjusted in order to decrease the number of unfeasible results for further analysis. Therefore, in order to obtain the classification results of this work, 0.5 was defined as the evaluation threshold, where the CUs with a model output greater than or equal to 0.5 were classified as CUs with NTLs, and CUs with output less than 0.5 were classified as regular.

To analyze the results of the classification model focusing on the identification of NTLs, error matrices were generated with the proposed applications. Table 11 presents the results of the error matrix for the model with temperature input, which has as a differential the input variables Pearson’s correlation and Euclidean distance, which reflect the use of the exogenous temperature variable. In Table 12, the results of the error matrix are presented for the model without these exogenous data entries, considering only the endogenous data of the distribution company.

Table 11. Error matrix—Results with temperature.

		Predicted Class	
		0	**1**
True Class	**0**	6879	1589
	1	1511	1171

Table 12. Error matrix—Results without temperature.

		Predicted Class	
		0	**1**
True Class	**0**	7043	1425
	1	1679	1003

Making the analysis by pairs, as observed in Tables 10 and 11, there was an improvement in the TP hits, which refers to the CUs with NTL. This improvement went from 1003 to 1171 CUs, a difference of 168 more CUs identified by the classification model, with the insertion of the temperature variable through the Pearson correlation and the Euclidean distance. However, the FP result, which represents the classification result of the regular CUs, worsened when inserting the exogenous variable, from 7043 to 6879 CUs. This may be related to the increase in electricity consumption during the winter, since temperatures are milder and consumption is higher in some CUs, as portrayed in [28]. Additionally, [20] pointed out similar consumer behaviors regarding the influence of temperature on energy consumption. Even though in [20], daily consumption data were available, by expanding the horizon of analysis, it was possible to make this observation.

In this same scenario, taking into account that the assessment for FN and FP should be minimized, inserting the exogenous variable improved the FN result, with a reduction from 1679 to 1511 CUs, and worsened the FP results, which went from 1425 to 1589 CUs classified in the model with the insertion of temperature. This worsening in the FP result, as in the case of TN, may also be related to thermal comfort issues during the greatest differences in annual thermal amplitude, evidenced mainly during winter.

Additionally, Table 13 presents the evaluation metrics for the model, considering the results obtained by evaluating the model with and without the insertion of the exogenous variable, temperature.

Table 13. Metrics-Results.

Metric	With Temperature	Without Temperature	%Dif
Accuracy	72.20%	72.16%	0.04%
Precision (1)	42.43%	41.31%	1.12%
Precision (0)	81.99%	80.75%	1.24%
Recall (1)	43.66%	37.40%	6.26%
Recall (0)	81.24%	83.17%	−1.94%
ROC-AUC	0.6849	0.6636	0.0213
PR-AUC	0.4168	0.3856	0.0312

As shown, the evaluation metrics had a significant improvement with the inclusion of temperature. There was an evolution of the model's accuracy, although not very expressive, accounting for an evolution of 0.04%. However, as mentioned, this metric should not be evaluated individually in classification models with an unbalanced database. Instead, looking at precision, precision improved by 1.12% and 1.24%, respectively, for CUs with NTLs and regular CUs.

Later, recall for identifying CUs with NTLs improved from 37.40% to 43.66%, an increase of 6.26%. However, this same metric declined for the identification of regular CUs, the only metric that decreased. Finally, the ROC-AUC and PR-AUC metrics also improved with the insertion of temperature, with 0.6636 to 0.6849 for the ROC-AUC, and from 0.3856 to 0.4168 for the PR-AUC. It is noteworthy that the identification of CUs with NTLs is one of the most relevant metrics for this work. Similar metric analyses and considerations are established by [7], even though in that case, temperature is not employed.

Additionally, in order to measure the developed model, the ROC curves were plotted in Figure 9, and good behavior of the training, test, and validation datasets can be observed. The ROC curves had an approximation, that is, the gap between the curves was low, and the separation of the curves may be a sign of model overfitting. In addition, the difference in ROC-AUC from 0.6636 to 0.6849 is another indication of the improvement of the model with the inclusion of temperature.

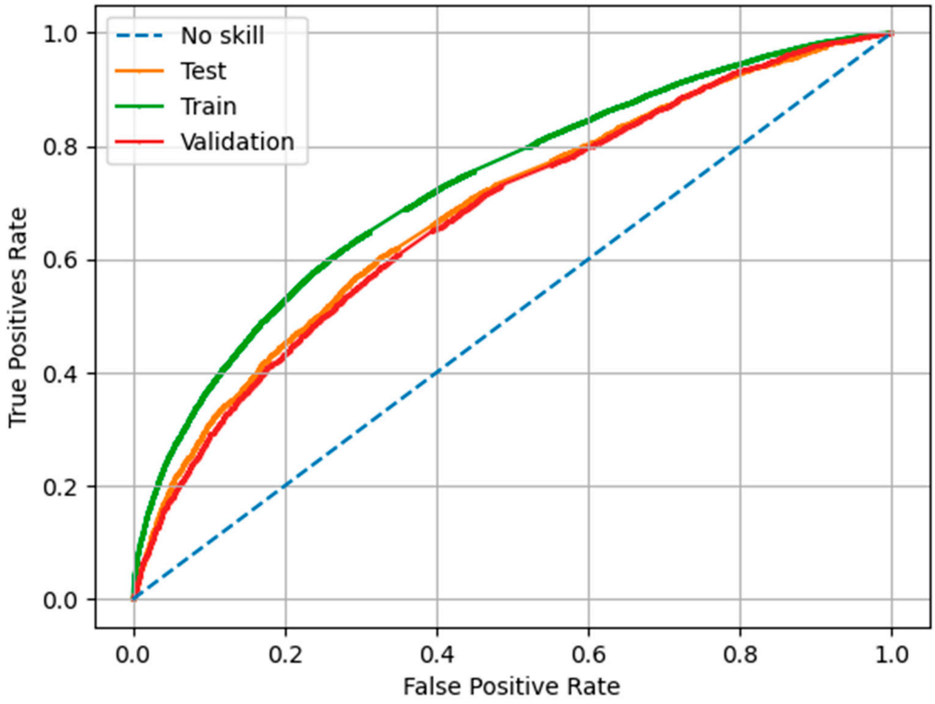

Figure 9. ROC-Curve—Results with Temperature.

The PR curve in Figure 10 summarizes the relationship between precision and recall, using the variations in the evaluation thresholds, as established in [7]. This result corroborates the difference between the PR-AUC of the model with temperature of 0.4168 and 0.3856 without temperature.

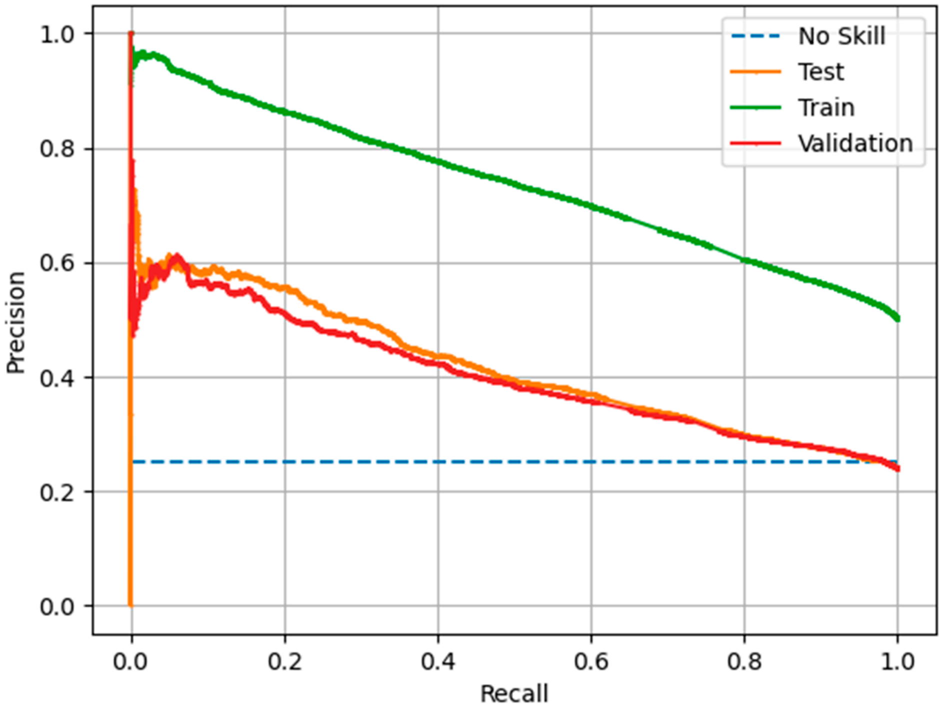

Figure 10. PR-Curve—Results with Temperature.

Rating Model Response Evaluation

After the training of the model was completed and the analysis of the model on a test basis was performed, the model was put to the test by observing the results of the models with and without the exogenous temperature variable, represented by Pearson correlation and Euclidean distance. Thus, from the granular results represented in the error matrix of Tables 10 and 11, two CUs were selected for the analysis of the consumption profiles of these units compared with temperature. It is noteworthy that the model acquires knowledge of the most varied patterns of consumption, and the pattern by which the model predicts consumers with NTLs and regular usage is difficult to verify. These CUs were selected following the principle of CU appearance as TP in both models.

Therefore, CU-1 represents a CU that was selected in both models, with the classification in the error matrices as TP. On the other hand, CU-2 was chosen because it was a CU that was identified as TP only in the model that had the exogenous variables with input; however, in the model without the Pearson correlation and Euclidean distance, this CU was selected as FN. From these definitions, the consumption profile of the CU-1 is presented in contrast to the information on the monthly average temperature for the same period, as shown in Figure 11.

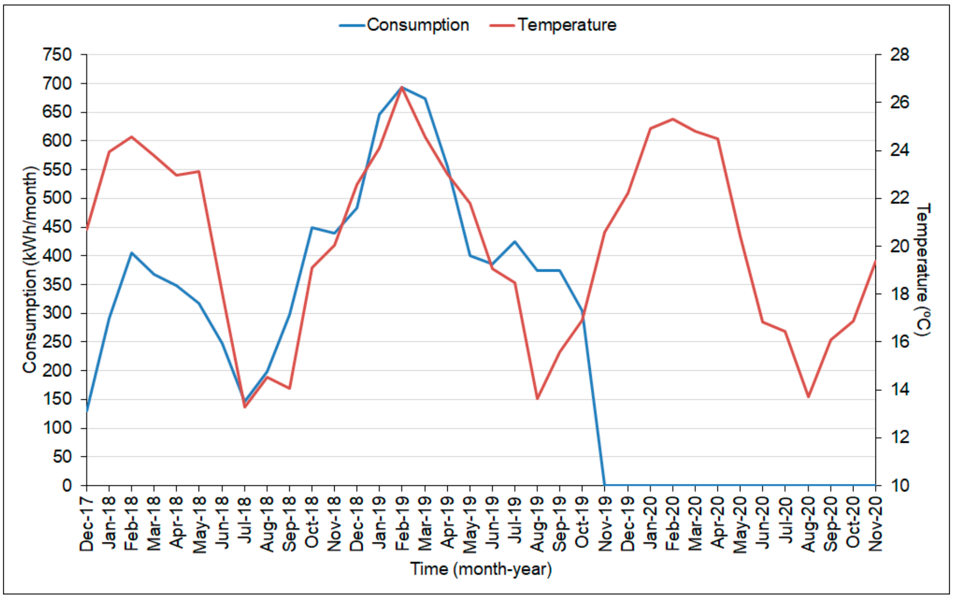

Figure 11. CU-1 Consumption Profile.

Figure 11 is a characteristic profile of a CU with measurement problems, where it was recorded in the field inspection as an equipment malfunction. It can be seen that as of November 2019, the consumption readings are all zero, signaling the beginning of the irregularity that resulted in the NTLs. However, prior to the start date of the irregularity, the consumption profile coincided in shape and pattern with the curve of monthly temperature averages, with similar peaks and valleys. The Pearson correlation and Euclidean distance for CU-1 were, respectively, 0.263 and 0.429. In terms of Pearson's correlation, the result shows a low correlation, which indicates probable outliers. As for the Euclidean distance, it was low but not located in the region of the greatest information gains for the identification of NTLs; as seen in Figure 8, it has little representation in the solution.

Then, the graphic result of CU-2 is shown in Figure 12, which represents a CU that was classified as NTL only in the model that had exogenous variables as additional inputs.

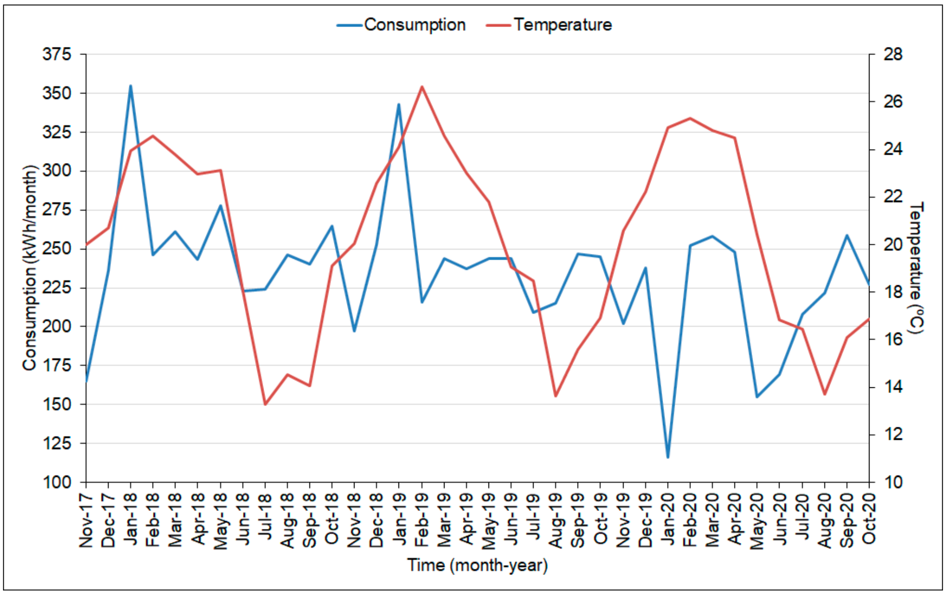

Figure 12. CU-2 Consumption Profile.

As shown in Figure 12, the consumption profiles and the average monthly temperature represent a certain pattern when evaluating the first half of the information contained in the graph, where it is possible to observe some almost coincidental peaks. However, the inconsistencies in consumption in the second half were mainly evident in the period from October 2019 to May 2020, which comprised the period of higher temperatures when there was a parallel reduction in consumption, as shown in the previous study by [31]. The Pearson correlation and Euclidean distance for this CU were 0.199 and 0.099, results that are close to CU-1, but with the result of the Euclidean distance in the region with the highest gain, facts that classify the CU with great chances of be committing a NTL.

Therefore, according to the result of the field inspection, CU-2 had consumption that was not being accounted for by the meter, which is reported as an energy deviation. Since this CU was identified as NTL only in the model with exogenous variables, this corroborated the irregularity of this CU, i.e., the energy deviation, which is often not easy to identify with models without exogenous information. In this way, it was possible to verify the importance of the exogenous temperature variable, which was associated with endogenous information to the energy distribution concessionaire, enabling the improvement of outlier classification models, focusing on the identification of CUs that are defrauding or stealing electricity.

Despite the difficulty in identifying consumption and temperature patterns by the model, even more so in the case of a deep learning model, in which the model weights for each input are difficult to understand, the model obtained good robustness in the field. Analyzing the current model, considering the field inspections in the identification of CU with NTL, robustness around 30% of the inspected targets was obtained.

4. Conclusions

The identification of NTL in energy CU is a challenge with numerous difficulties. It has significant technical difficulty that must be mitigated by energy companies, with both data-oriented methods and other available resources. It is noteworthy that it requires an investment in capital to move technical teams to the places where they can be identified and normalized. In view of this, this work proposed a data-oriented method for CU identification with NTLs.

The work validated the use of exogenous data from the climate database to improve the robustness of the ML model of CU classification with NTLs. Furthermore, with the use of temperature, it was possible to add value to the distribution company data and achieve metrics on a test basis with better robustness and with good applicability in the field for the distribution concessionaire. In particular, robustness increased for the model with the exogenous data inputs.

In evaluating the results metrics, the ROC-AUC with a result of 0.6845 and the CU recall (1) with a NTL of 43.66% in the test bench stand out. These metrics presented the best improvement with the introduction of temperature time series to the model. That highlights that the framework developed in this research contributed to improving identifying NTL outliers. Additionally, with the detailed evaluation of CU signaled by the model with exogenous variables, the importance of this variable in the identification of NTLs is reaffirmed. Finally, the test of the methodology was started in the field, and robustness around 30% was obtained in the first tests. As a continuation of this work, the direction should focus on the aggregation of other exogenous data through web scraping aiming at making the methodology more robust.

Author Contributions: M.B.C., B.K.H., R.G.N., F.G.K.G., L.R.P., N.K.N. and A.d.R.A. contributed equally to the idealization of this paper, elaboration of the methodology, and conception, as well in the analysis of the results and the writing and reviewing of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: The authors are grateful for the technical and financial support from the State Electric Energy Company and the Equatorial Energia Group (ANEEL R&D Program through the CEEE/EQUATORIAL/UFSM project no. 5000003849), National Institute of Science and Technology in Distributed Generation Systems (INCTGD), National Council for Scientific and Technological Development (CNPq—no. 465640/2014-1), Coordination for the Improvement of Higher Education Personnel (CAPES—no. 23038.000776/2017-54), Research Support Foundation of the State of Rio Grande do South (FAPERGS—no. 17/2551-0000517-1) and Federal University of Santa Maria (UFSM), Brazilian Institutions.

Data Availability Statement: The data presented in this study might be available on request from the corresponding author. The data are not publicly available due to confidentiality agreement with Equatorial Energia Group.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chandel, P.; Thakur, T.; Sawale, B.A. Energy Meter Tampering: Major Cause of Non-Technical Losses in Indian Distribution Sector. In Proceedings of the 2016 International Conference on Electrical Power and Energy Systems (ICEPES), Bhopal, India, 14–16 December 2016; pp. 368–371.
2. Carr, D.; Thomson, M. Non-Technical Electricity Losses. *Energies* **2022**, *15*, 2218. [CrossRef]
3. Coma-Puig, B.; Carmona, J. Bridging the Gap between Energy Consumption and Distribution through Non-Technical Loss Detection. *Energies* **2019**, *12*, 1748. [CrossRef]
4. Haykin, S. Redes Neurais. Available online: https://integrada.minhabiblioteca.com.br/#/books/9788577800865/ (accessed on 20 June 2021).
5. Li, Q.; Yan, M.; Xu, J. Optimizing Convolutional Neural Network Performance by Mitigating Underfitting and Overfitting. In Proceedings of the 2021 IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS), Shanghai, China, 23–25 June 2021; pp. 126–131.
6. Chatterjee, S.; Archana, V.; Suresh, K.; Saha, R.; Gupta, R.; Doshi, F. Detection of Non-Technical Losses Using Advanced Metering Infrastructure and Deep Recurrent Neural Networks. In Proceedings of the 2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I CPS Europe), Milan, Italy, 6–9 June 2017; pp. 1–6.
7. Buzau, M.-M.; Tejedor-Aguilera, J.; Cruz-Romero, P.; Gómez-Expósito, A. Hybrid Deep Neural Networks for Detection of Non-Technical Losses in Electricity Smart Meters. *IEEE Trans. Power Syst.* **2020**, *35*, 1254–1263. [CrossRef]
8. Hasan, M.d.N.; Toma, R.N.; Nahid, A.-A.; Islam, M.M.M.; Kim, J.-M. Electricity Theft Detection in Smart Grid Systems: A CNN-LSTM Based Approach. *Energies* **2019**, *12*, 3310. [CrossRef]
9. Kabir, B.; Shams, P.; Ullah, A.; Munawar, S.; Asif, M.; Javaid, N. Detection of Non-Technical Losses Using MLP-GRU Based Neural Network to Secure Smart Grids. In Proceedings of the Conference on Complex, Intelligent, and Software Intensive Systems, Asan, Korea, 1–3 July 2021.

10. Dong, L.; Li, Q.; Wu, K.; Fei, K.; Liu, C.; Wang, N.; Yang, J.; Li, Y. Nontechnical Loss Detection of Electricity Based on Neural Architecture Search in Distribution Power Networks. In Proceedings of the 2020 International Conference on Smart Grid and Clean Energy Technologies (ICSGCE), Sarawak, Malaysia, 4–7 October 2020; pp. 143–148.
11. CASTRO, D.G.F.L.N.D. Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações. Available online: https://integrada.minhabiblioteca.com.br/#/books/978-85-472-0100-5/ (accessed on 20 June 2021).
12. Ghori, K.M.; Awais, M.; Khattak, A.S.; Imran, M.; Fazal-E-Amin; Szathmary, L. Treating Class Imbalance in Non-Technical Loss Detection: An Exploratory Analysis of a Real Dataset. *IEEE Access* **2021**, *9*, 98928–98938. [CrossRef]
13. Rodríguez, F.; Lecumberry, F.; Fernández, A. Non Technical Loses Detection-Experts Labels vs. Inspection Labels in the Learning Stage. In Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods—ICPRAM; SciTePress: Setúbal, Portugal, 2014; pp. 624–628.
14. Markoč, Z.; Hlupić, N.; Basch, D. Detection of Suspicious Patterns of Energy Consumption Using Neural Network Trained by Generated Samples. In Proceedings of the ITI 2011, 33rd International Conference on Information Technology Interfaces, Cavtat/Dubrovnik, Croatia, 7–30 June 2011; pp. 551–556.
15. Andrew, N.G. The AI Pioneer Says It's Time for Smart-Sized, "Data-Centric" Solutions to Big Issues. Available online: https://spectrum.ieee.org/andrew-ng-data-centric-ai (accessed on 22 February 2022).
16. Angelos, E.W.S.; Saavedra, O.R.; Cortés, O.A.C.; de Souza, A.N. Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems. *IEEE Trans. Power Deliv.* **2011**, *26*, 2436–2442. [CrossRef]
17. Guerrero, J.I.; Monedero, I.; Biscarri, F.; Biscarri, J.; Millán, R.; León, C. Non-Technical Losses Reduction by Improving the Inspections Accuracy in a Power Utility. *IEEE Trans. Power Syst.* **2018**, *33*, 1209–1218. [CrossRef]
18. Madrigal, M.; Rico, J.J.; Uzcategui, L. Estimation of Non-Technical Energy Losses in Electrical Distribution Systems. *IEEE Lat. Am. Trans.* **2017**, *15*, 1447–1452. [CrossRef]
19. Simões, P.F.M.; Souza, R.C.; Calili, R.F.; Pessanha, J.F.M. Analysis and Short-Term Predictions of Non-Technical Loss of Electric Power Based on Mixed Effects Models. *Socio-Econ. Plan. Sci.* **2020**, *71*, 100804. [CrossRef]
20. Hu, W.; Yang, Y.; Wang, J.; Huang, X.; Cheng, Z. Understanding Electricity-Theft Behavior via Multi-Source Data. In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 2264–2274. [CrossRef]
21. Aslam, Z.; Javaid, N.; Ahmad, A.; Ahmed, A.; Gulfam, S.M. A Combined Deep Learning and Ensemble Learning Methodology to Avoid Electricity Theft in Smart Grids. *Energies* **2020**, *13*, 5599. [CrossRef]
22. Christen, R.; Mazzola, L.; Denzler, A.; Portmann, E. Exogenous Data for Load Forecasting: A Review. In *Proceedings of the 12th International Joint Conference on Computational Intelligence—Volume 1: CI4EM*; SciTePress: Setúbal, Portugal, 2020; pp. 489–500. [CrossRef]
23. Pérez-Ortiz, M.; Jiménez-Fernández, S.; Gutiérrez, P.A.; Alexandre, E.; Hervás-Martínez, C.; Salcedo-Sanz, S. A Review of Classification Problems and Algorithms in Renewable Energy Applications. *Energies* **2016**, *9*, 607. [CrossRef]
24. Cybenko, G. Approximation by Superpositions of a Sigmoidal Function. *Math. Control. Signals Syst.* **1989**, *2*, 303–314. [CrossRef]
25. Haben, S.; Giasemidis, G.; Ziel, F.; Arora, S. Short Term Load Forecasting and the Effect of Temperature at the Low Voltage Level. *Int. J. Forecast.* **2019**, *35*, 1469–1484. [CrossRef]
26. Hu, L.; Zhang, L.; Wang, T.; Li, K. Short-Term Load Forecasting Based on Support Vector Regression Considering Cooling Load in Summer. In Proceedings of the 2020 Chinese Control and Decision Conference (CCDC), Hefei, China, 22–24 August 2020; pp. 5495–5498.
27. Hobby, J.; Tucci, G. Analysis of the Residential, Commercial and Industrial Electricity Consumption. In Proceedings of the Innovative Smart Grid Technologies Asia (ISGT), Perth, Australia, 13–16 November 2011; pp. 1–7.
28. Silva, L.N.; Abaide, A.R.; Negri, V.G.; Capeletti, M.; Lopes, L.F.; Cardoso, G. Diagnostic and Input Selection Tool Applied on Weather Variables for Studies of Short-Term Load Forecasting. In Proceedings of the 2019 8th International Conference on Modern Power Systems, MPS 2019, Cluj Napoca, Romania, 21–23 May 2019. [CrossRef]
29. Yurtseven, Ç. The Causes of Electricity Theft: An Econometric Analysis of the Case of Turkey. *Util Policy* **2015**, *37*, 70–78. [CrossRef]
30. Yuejun, H.; Fubin, L.; Jieqing, X.; Tingting, M. Non-Technical Loss Detection by Multi-Dimensional Outlier Analysis on the Remote Metering Data. In Proceedings of the 2016 China International Conference on Electricity Distribution (CICED), Xi'an, China, 10–13 August 2016; p. 2545.
31. Capeletti, M.B.; da Rosa Abaide, A.; Hammerschmitt, B.K.; Neto, N.K.; dos Santos, L.L.C.; Milbradt, R.G.; Guarda, F.G.K.; Prade, L.R.; da Rosa Moreira, G. Descriptive Data Analysis of Weather Inputs for Non-Technical Losses Detection System. In Proceedings of the 2021 9th International Conference on Modern Power Systems (MPS), Cluj-Napoca, Romania, 16–17 June 2021; p. 623.
32. INMET (Instituto Nacional de Meteorologia). Available online: https://portal.inmet.gov.br/dadoshistoricos (accessed on 8 February 2022).
33. Iglesias, F.; Kastner, W. Analysis of Similarity Measures in Times Series Clustering for the Discovery of Building Energy Patterns. *Energies* **2013**, *6*, 579–597. [CrossRef]
34. Jawad, M.; Nadeem, M.; Shim, S.-O.; Khan, I.; Shaheen, A.; Habib, N.; Hussain, L.; Aziz, W. Machine Learning Based Cost Effective Electricity Load Forecasting Model Using Correlated Meteorological Parameters. *IEEE Access* **2020**, *8*, 146847–146864. [CrossRef]

35. Zhi, X.; Yuexin, S.; Jin, M.; Lujie, Z.; Zijian, D. Research on the Pearson Correlation Coefficient Evaluation Method of Analog Signal in the Process of Unit Peak Load Regulation. In Proceedings of the 2017 13th IEEE International Conference on Electronic Measurement and Instruments (ICEMI), Yangzhou, China, 20–22 October 2017; pp. 522–527.
36. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow:Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2016**. [CrossRef]
37. Kingma, D.P.; Ba, J. ADAM: A method for stochastic optimization. In Proceedings of the 3rd International Conference for Learning Representations—ICLR 2015, San Diego, CA, USA, 7–9 May 2015. [CrossRef]
38. Pereira, J.; Saraiva, F. A Comparative Analysis of Unbalanced Data Handling Techniques for Machine Learning Algorithms to Electricity Theft Detection. In Proceedings of the 2020 IEEE Congress on Evolutionary Computation (CEC), Glasgow, UK, 19–24 July 2020; pp. 1–8.
39. Huang, J.; Ling, C.X. Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 299–310. [CrossRef]
40. Da Silva, L.A.; Peres, S.M.; Boscarioli, C. Introdução à Mineração de Dados-Com Aplicações Em R. Available online: https://integrada.minhabiblioteca.com.br/#/books/9788595155473/ (accessed on 20 September 2021).
41. Buzau, M.M.; Tejedor-Aguilera, J.; Cruz-Romero, P.; Gómez-Expósito, A. Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning. *IEEE Trans Smart Grid* **2019**, *10*, 2661–2670. [CrossRef]
42. Witten, I.H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. *SIGMOD Rec.* **2002**, *31*, 76–77. [CrossRef]