International Journal of Advanced Engineering and Management Research

Vol. 8, No. 05; 2023

ISSN: 2456-3676

Electrical Network Fraud Detection Using a CNN+LSTM Model

Rafael A. Gomez[1] and Felipe A. Llaugel[2]
[1]Instituto Tecnológico de Santo Domingo, Santo Domingo, Dominican Republic.
[2]Faculty of Engineering, Universidad Dominicana O&M, Santo Domingo, Dominican Republic.

doi: 10.51505/ijaemr.2023.8505 URL: http://dx.doi.org/10.51505/ijaemr.2023.8505

Received: Aug 31, 2023 Accepted: Sep 11, 2023 Online Published: Sep 15, 2023

Abstract
This work focuses on the persistent problem of energy loss in the Dominican Republic and in Latin America, with an emphasis on non-technical losses caused by electricity fraud. Given the seriousness of the problem, this paper proposes the development and implementation of a fraud classification model in electricity distribution networks using Deep Learning, specifically a combined model of a Convolutional Neural Network Distributed in Time (CNN) and Long Short Term Memory (LSTM). The goal is to understand and evaluate current fraud detection techniques, investigate the applicability and efficiency of CNN + LSTM models in fraud detection, and address potential challenges in implementing this model in Latin America. The justification lies in the considerable financial losses generated by electricity fraud.

Keywords: Energy loss; non-technical loss; energy fraud; neural networks; convolutional neural networks; recurrent neural networks; artificial intelligence

1. Introduction

Energy loss is one of the main issues in the electrical industry. This occurs worldwide to varying degrees, however, the Dominican Republic is among the Latin American countries where these losses occur most significantly. According to performance reports from the Dominican Corporation of State Electric Companies (CDEEE), the average loss among all distributors has been 30.4%. Electrical loss has two roots: technical loss and non-technical loss.

Technical loss is caused by inherent reasons related to the generation and distribution of electricity. A part of this is fixed loss, which naturally occurs in energy creation, via the heating of transformer cores, the corona effect on transmission lines, and others. There is also variable loss, which is related to energy transportation. These losses are controlled through technical interventions on the components of the electrical network, which reduce them.

Non-technical loss represents the largest proportion of loss in the Dominican Republic, and it is caused by factors external to the electrical systems, related to the economy and the management of energy companies. Two of the most significant areas are theft and fraud: citizens illegally connect to a power source or manipulate consumption measurement, resulting in a reduction in their bill. On the distributor's side, a bill is generated that is lower than the consumption, causing a significant deficit.

Vol. 8, No. 05; 2023

ISSN: 2456-3676

Research and statistical models have been carried out using machine learning technologies with the aim of classifying fraudulent users in real time, in order to recover this loss. The application of science has evolved over the past decade, moving from Game Theory and state estimation methods to more sophisticated strategies, such as combining different neural network architectures for classification, clustering for feature detection, and subsequent detection of fraudulent users through Random Forest algorithms, up to the applications of KNN, XG-Boost, and CNN for the same purpose.

This study provides a tool for electric fraud discovering and inspection resources optimization.

2. Problem Statement

2.1 General Objective

The general objective of this research project is to develop a classification model using Deep Learning techniques, specifically a model of Time-Distributed Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), for fraud detection in Electric Power Distribution Networks.

2.2 Research Questions

- Can electric fraud in energy distribution networks be accurately detected using Time Distributed CNN + LSTM models?
- What is the potential impact and savings for the public and private sector in the Dominican Republic by applying these models for fraud detection?
- What are the specific challenges and limitations in implementing this model, particularly in terms of data quality and imbalance?

2.3 Justification

Fraud in the distribution of electrical energy has significant implications for both the public and private sectors within Latin America, generating substantial financial losses. The detection of these non-technical losses has gained attention in recent decades, however, the application of advanced deep learning techniques such as Time Distributed CNN + LSTM for this purpose within the Latin American context remains limited. Electricity Distribution Companies (EDC) have not managed to meet the goals established in the National Pact for the Reform of the Electricity Sector regarding the reduction of energy losses. This non-compliance is evident in the performance reports of the electricity sector, where it is shown that loss rates have not reached the ranges established in the Electric Pact's regulations. Given this situation, there is a justified need to carry out research on electrical fraud detection using neural networks. Through the use of these networks, energy consumption could be analyzed in real time and suspicious patterns could be detected that may indicate fraudulent activities.

Vol. 8, No. 05; 2023

ISSN: 2456-3676

Table 1: Relevant values for public action on the subject of energy loss from the main distribution companies. Source: Unified Council of Electricity Distribution Companies (CUED)

Energy Loss Management Indicators of main distribution companies in Dominican Republic					
Distribution Companies	Projection 2021	Actual 2021	Projection 2022	Actual 2022	As of date 2023
EDENORTE	20.5%	21.9%	18.5%	21.7%	21.7%
EDESUR	21.4%	26.2%	19.4%	26.5%	26.2%
EDEESTE	38.3%	48.4%	32.3%	47.6%	44.6%
PROMEDIO	26.7%	32.2%	23.4%	31.9%	30.8%

Previous studies have achieved promising results in fraud detection through these models, with unique benefits including the ability to utilize both reliable and unreliable consumption data sources, and incremental retraining capabilities for model improvement. Given these potential advantages and the relevance of addressing fraud in the distribution of electrical energy, this research is crucial for advancing the body of knowledge within this region.

The results of this research have a wide scope, with potential benefits for the public and private sectors of Latin America. By detecting fraudulent activities, significant savings could be achieved, supporting the economic development and stability of the region. Additionally, the successful application of the proposed model could generate broader implications in the context of machine learning applications for fraud detection and could serve as a benchmark for future research aimed at optimizing and improving these techniques. The dataset used in model construction was provided by EDEESTE (one of the 3 electricity distribution companies), and consisted of daily observations of electric meters from April 2022 to April 2023.

3. Literature Review

Public service companies, such as electricity providers, play a vital and transformative role in modern society. The service they offer is fundamental for the development of everyday life, and therefore its proper administration and resource optimization has emerged as a key area of interest in scientific and technical research. One focus of this optimization is the minimization of non-technical loss, which is attributed to losses incurred by the distributor due to fraudulent practices or metering failures.

At the beginning of the century, Rong Jiang et al. (2002) proposed a classification method that combined multiple statistical techniques, including the notable wavelet method. This decision to employ wavelets in the classification process, as highlighted in their work, was adopted due to the unique ability of this technique to capture localized features, leading to a model with higher accuracy than the prevailing methodologies of its time. The input data for their classifier consisted of an aggregate of consumption data from meters installed on consumers properties,

Vol. 8, No. 05; 2023

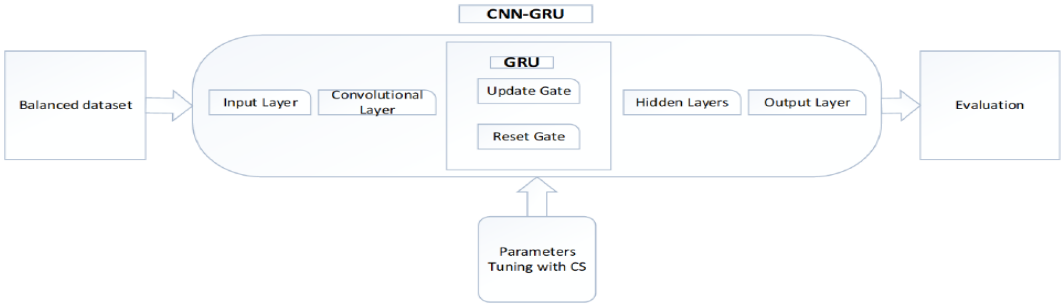
ISSN: 2456-3676

labeled as "honest" or "fraudulent". At this point, it is important to emphasize the demonstrated correlation between the amount of sample data used for training and the generalization capacity of the resulting model.

With the advent and rapid evolution of neural networks, with their widely recognized and powerful classification capabilities, multiple iterations have been made that combine various statistical and mathematical methods to solve this problem. This research has gained traction as the need for an effective solution has been increasing. The most influential contributions in this area have focused on both supervised and unsupervised techniques, leveraging a variety of models and architectures that have stimulated advancement in the field.

In this context, a study carried out by a research team from the Polytechnic University of Catalonia, led by Coma-Puig (2016), stood out for its supervised Machine Learning approach. They used a dataset similar to previous studies, but incorporated a unique nuance by including multiple energy sources, such as gas, and other descriptive variables about the meters. These additional variables included factors such as the tariff, the type of climate, and gas consumption instead of electricity. The model they proposed also had the distinctive feature of being online, that is, capable of integrating and adapting to the results of emerging fraudulent dynamics.

In a more recent study, Nasir Ayub et al. (2022) have proposed a new approach to the issue, with the study "Predictive Data Analytics for Electricity Fraud Detection Using Tuned CNN Ensembler in Smart Grid". In this ambitious research, the authors developed a fraud detection model that merges a Convolutional Neural Network (CNN) with a Gated Recurrent Unit (GRU). In addition, the tuning of the hyperparameters of the proposed model is carried out through a metaheuristic method called Cuckoo Search. They have also addressed the problem of class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE). The results of simulations based on real energy consumption data show that the proposed model (CNN-GRU-CS) outperforms other methodologies in terms of effectiveness and accuracy, with an average increase of 10%. The calculated accuracy of the proposed method is 92%, and the precision is 94%, showing a significant advance in the field of electrical fraud detection.



Vol. 8, No. 05; 2023

ISSN: 2456-3676

Figure 1. Graphical representation of the CNN-GRU-CS architecture used in the study "Predictive Data Analytics for Electricity Fraud Detection Using Tuned CNN Ensembler in Smart Grid".

It is important to emphasize that despite the recent focus on neural networks, efforts have also been made with simpler architectures that have had excellent results, such as the support vector models architectures presented by Petrlik et al. (2022) and decision tree architectures presented by Appiah et al. (2023) in the study "Extremely randomized trees machine learning model for electricity theft detection." In the latter, the use of the grid search optimization technique to optimize the hyperparameters of the proposed model stands out, as well as the fact of very promising metric yields (98% accuracy, 98% F1, 95% Matthew correlation.)

There are also fascinating studies based on unsupervised learning, such as the one proposed by Dai et al. (2022), this model was a Variational Recurrent Autoencoder with attention. In this paper, titled "Smart Meter Data Anomaly Detection using Variational Recurrent Autoencoders with Attention" the researchers express that this model adapts very well to a permanent problem in the research efforts of this topic: the data received are usually complex to deal with, with many null values in the fraudulent cases. This method can successfully detect different types of anomalies, including global and contextual.

Table 2. Variational Auto encoder hyperparameters proposed by Dai et al. Source: Smart Meter Data Anomaly Detection using Variational Recurrent Autoencoders with Attention

Proposed model hyperparameters	
LSTM hidden layers	2
Units in hidden layers	218
Sequence length (W)	168
Latent dimensions	3
Training iterations	550
Learning rate	0.0001
Batch size	1024
Optimizer	Adam
Optimizer Learning Rate	0.01

Table 3. Performance of the proposed model discussed above. Source: Smart Meter Data Anomaly Detection using Variational Recurrent Autoencoders with Attention

Rendimiento en conjunto de datos sintétitco			
Method	Metricas		
	P	*R*	*F1*
CBLOF	0.65	0.68	0.66
KNN	0.69	0.69	0.69
PCA	0.83	0.84	0.83
OCSVM	0.82	0.82	0.82
VAE-baseline	0.89	0.9	0.9
Ours	**0.95**	**0.93**	**0.94**

Vol. 8, No. 05; 2023

ISSN: 2456-3676

Autoencoders have proven to be a constant in these advancements due to their powerful ability to learn and reconstruct the normal patterns of a statistical distribution. Chahla et al. (2019) combined this architecture with LSTM and K-means for the problem presented, also obtaining very good results. Shortly before, researchers Joao Pereira and Margarida Silveira (2018) had presented a similar proposal, in which a Variational Recurrent Autoencoder with attention was used.

The field of electricity fraud detection has experienced significant advancements thanks to the application and development of machine and deep learning techniques. This opens a wide field for future research, where new strategies and approaches can be explored and developed to further improve the efficiency and accuracy of fraud detection. On the other hand, these advancements represent the resolution of a costly problem for the Latin American region. The motivation for this work lies in contributing to this collective effort, always seeking innovative solutions to optimize energy consumption and minimize fraud.

4. Methodology used

The dataset used in this study was obtained from one of the main electricity distributors in the Dominican Republic and consists of daily consumption readings (in kWh) from approximately 24,000 smart meters during the period between April 1, 2022, and March 31, 2023. The dataset included readings from 20,000 non-fraudulent smart meters and 4,000 fraudulent ones.

4.1 Data Preprocessing stage

The data preprocessing stage is essential to convert the raw EC data obtained into a useful data structure that our model can train on and use to generate predictions.

4.1.1 Data Reduction

As with any dataset, there are traces of null values. In this case, there were dates for which certain meters did not emit data. Cases where a meter had null values in more than fifty percent were eliminated. Cases were also removed where the reading was constant throughout the period. It was deduced that there was no consumption at all, so it was not relevant. Finally, duplicates were also eliminated.

4.1.2 Feature Extraction, Interpolation, and Normalization

Linear interpolation was performed to estimate the reading on the days when no signal was received from the meter. After this, daily consumption was calculated by looking for the difference in readings between consecutive days. Data normalization ensures that the meter data are not biased, as machine learning models are sensitive to this condition. We used MIN-MAX normalization to scale the data in a form that the proposed model can easily use. Data normalization is achieved according to the equation below:

$$f(x_i)=(x_i-\min(x))/(\max(x)-\min(x))$$

where x_i represents the value on a day for a meter, and $\min(x)$ and $\max(x)$ represent the minimum and maximum values for that meter in the time series.

Vol. 8, No. 05; 2023

ISSN: 2456-3676

We concatenate a descriptor that represents the number of "blinkouts" or instances when data transmission was unsuccessful.

4.2 Handling of Unbalanced Data Set

A common inconvenience evident in the literature review on addressing this problem is the recurrence of an unbalanced dataset. This is due to the fact that the number of meters reported as fraudulent is a tiny fraction of the total number of meters. To prevent this, data providers were asked to identify and include as many irregular meters as possible. After carrying out the preprocessing explained above, the data matrix of fraudulent meters was around 700. To balance the data, we considered the dilemma of selecting 700 regular meters or performing oversampling on the irregular ones. Since in the iterations performed, the models tended to have better generalization power, we decided to do the oversampling, increasing the irregular ones from 700 to 4,000.

4.3 CNN + LSTM Model Training

The deep learning model implemented in this study was a sequential combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory units (LSTM). The model architecture began with a time-distributed wrapper layer encapsulating a 1D convolutional layer with 64 filters and a kernel size of 3, followed by a max pooling operation. Then, the model flattened these features to feed the LSTM layer with 50 units. The output layer consisted of a single unit with a sigmoid activation function to predict the binary outcome. The model was compiled with a learning rate of 0.01 using the Adam optimizer and the binary cross-entropy loss function. The model was trained on a reshaped version of the balanced training dataset with each training epoch divided into a 20% validation set. Dropout was initially considered as a regularization technique, but it was observed that the model performed better without it.

4.4 Model Evaluation

The metrics chosen to evaluate the model were: Accuracy, Precision, Recall, and F1. As part of the model's hyperparameter optimization process, we took into account the fact that False Positives should be minimized, as this would imply including them in a regularization campaign on the part of the distributing company, and by reducing them, the company's resources would be optimized.

(1)	$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$
(2)	$Precision = \frac{TP}{TP + FP}$
(3)	$Recall = \frac{TP}{TP + FN}$

Vol. 8, No. 05; 2023

ISSN: 2456-3676

$$F1 = \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

5. Results

The model's results, compared with related efforts, are displayed next:

Table 4. Comparison of proposed model with related works.

Model	Country	Accuracy	AUC	Precision	Recall	F1-Score
Wide and Deep CNN (Zheng et al., 2018)	China	N/A	0.78	N/A	N/A	N/A
CNN-LSTM (Hasan et al., 2019)	China	0.89	N/A	0.92	0.96	0.94
LSTM + bat-based RUS Boost (Adil et al., 2020)	China	0.87	0.87	0.88	0.91	N/A
Hybrid DNN (Buzau et al., 2020)	China	N/A	0.82	0.466	N/A	N/A
SVM (Anwar et al., 2021)	China	N/A	0.89	0.85	0.86	0.87
SSDA (Huang & Xu, 2021)	China	N/A	N/A	N/A	0.9	N/A
LSTM + GRU (Pamir et al., 2022)	China	0.91	0.91	0.97	0.86	0.91
AlexNet + AdaBoost + Artificial Bee colony (Ullah et al., 2022)	China	0.88	0.91	0.86	0.84	0.85
Proposed model	**Dominican Republic**	**0.96**	**0.98**	**0.95**	**0.98**	**0.96**

6. Conclusion

In this study, we addressed the problem of fraud detection in electricity distribution networks in Latin America, with a specific focus on the Dominican Republic. The high levels of energy loss in the country, mainly due to non-technical losses such as theft and fraud, have significant financial implications for both the public and private sectors. We proposed and implemented a classification model using deep learning techniques, specifically a model of Time-Distributed Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM), to detect fraudulent activities in the energy distribution sector.

The results showed promising outcomes in fraud detection, with a focus on minimizing false positives to optimize the resources of distribution companies. By leveraging reliable and unreliable consumption data and the model's capacity for incremental retraining, our approach demonstrated unique benefits.

This research contributes to the existing body of knowledge by applying advanced deep learning techniques to address the problem of fraud in electricity distribution networks in Latin America.

Vol. 8, No. 05; 2023

ISSN: 2456-3676

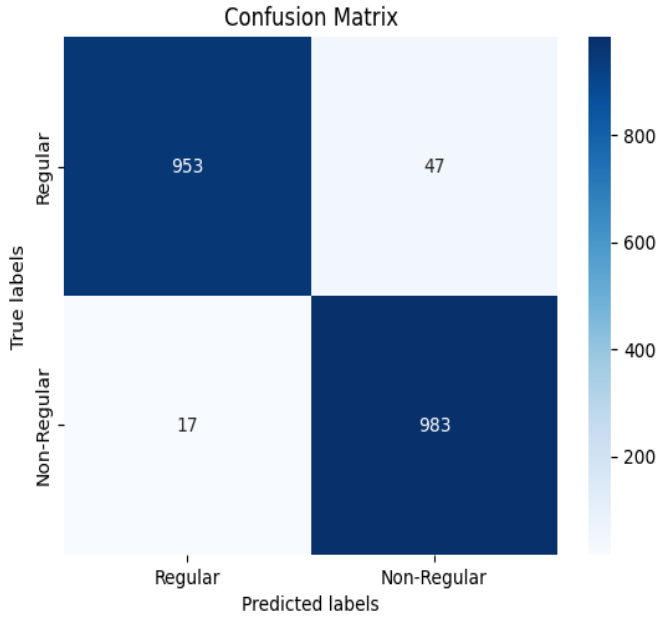
The study highlights the potential of using time-distributed CNN + LSTM models and provides insight into the challenges and limitations of implementing such models in the region.

The findings of this study have practical implications for electricity distribution companies in the Dominican Republic and other countries facing similar challenges. By implementing the proposed model, these companies can analyze energy consumption in real-time and identify suspicious patterns that may indicate fraudulent activities. This proactive approach can help reduce non-technical losses and improve the financial performance of the distribution sector.

Although low income families receive a subsidy to pay electric bill, this have not been an incentive for fraud reduction.

7. Appendices

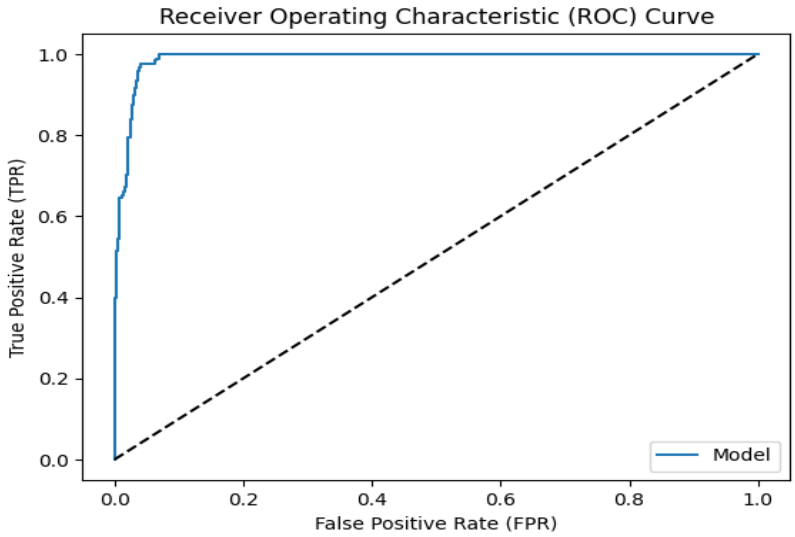
Apendix 1. Confusion matrix of the model evaluation results



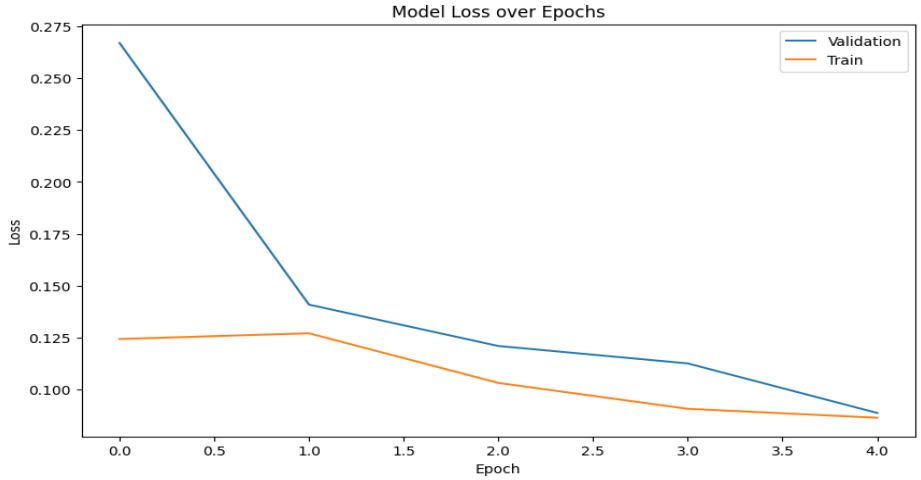
Appendix 2. ROC-AUC Graph

Vol. 8, No. 05; 2023

ISSN: 2456-3676



Appendix 3. Training and validation set loss history graph across epochs



References

Appiah, Stanley & Akowuah, Emmanuel & Ikpo, Valentine & Dede, Albert. (2023). Extremely randomised trees machine learning model for electricity theft detection. Machine Learning with Applications. 12. 100458. 10.1016/j.mlwa.2023.100458

Ayub N, Ali U, Mustafa K, Mohsin SM, Aslam S. (2022). Predictive Data Analytics for Electricity Fraud Detection Using Tuned CNN Ensembler in Smart Grid. Forecasting. ; 4(4):936-948. https://doi.org/10.3390/forecast4040051

B. Coma-Puig, J. Carmona, R. Gavaldà, S. Alcoverro and V. Martin, (2016).Fraud Detection in Energy Consumption: A Supervised Approach, 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Montreal, QC, Canada, 2016, pp. 120-129, doi: 10.1109/DSAA.2016.19

Vol. 8, No. 05; 2023

ISSN: 2456-3676

Chahla, Charbel & Snoussi, Hichem & Merghem, L. & Esseghir, Moez. (2019). A Novel Approach for Anomaly Detection in Power Consumption Data. Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods ICPRAM - Volume 1, 483-490, 2019, Prague, Czech Republic.

Dai, W., Liu, X., Heller, A., & Nielsen, P. S. (2021, October). Smart meter data anomaly detection using variational recurrent autoencoders with attention. In International Conference on Intelligent Technologies and Applications (pp. 311-324). Cham: Springer International Publishing.

DOI: 10.5220/0007361704830490

Emadaleslami, M., & Haghifam, M. (2021). A Machine Learning Approach to Detect Energy Fraud in Smart Distribution Network. International Journal of Smart Electrical Engineering, 10(02), 59-66. doi: 10.30495/ijsee.2021.683248

Jiang, R., Tagaris, H., Lachsz, A., & Jeffrey, M. (2002, October). Wavelet based feature extraction and multiple classifiers for electricity fraud detection. In IEEE/PES Transmission and Distribution Conference and Exhibition (Vol. 3, pp. 2251-2256). IEEE.

Pereira, J., & Silveira, M. (2018, December). Unsupervised anomaly detection in energy time series data using variational recurrent autoencoders with attention. In 2018 17th IEEE International Conference on machine Learning and Applications (ICMLA) (pp. 1275-1282). IEEE. 10.1109/ICMLA.2018.00207

Petrlik, I., Lezama, P., Rodriguez, C., Inquilla, R., Reyna-González, J. E., & Esparza, R. (2022). Electricity Theft Detection using Machine Learning. International Journal of Advanced Computer Science and Applications, 13(12)., 13(12), 2022.