

# Application of Time Series Clustering for Improving Forecasts in Energy Markets

Ines Araujo<sup>1</sup>, Rita Teixeira<sup>1</sup>, John Peñaloza Morán<sup>2</sup>, Tiago Pinto<sup>1,2</sup>, Jose Baptista<sup>1,2</sup>

<sup>1</sup>Universidade de Trás-os-Montes e Alto Douro, Vila Real, Portugal

<sup>2</sup>INESC TEC, Porto, Portugal

**Abstract**—The increasing integration of distributed energy generation into the electrical grid has led to changes in the structure and organization of energy markets over the past years. Market trading has become increasingly demanding due to the different types of production profiles. A forecast of the total production of all assets is made to bid for energy. Whenever there are differences between the forecast and the actual produced energy, a deviation occurs, which is assigned to the agent responsible for its settlement. This article proposes the application of a linear regression algorithm supported by a clustering method to forecast energy production. Based on the historical production profile of the installations in each cluster, it is possible to predict the production pattern for a period with no available data, thus standardizing this data for other assets belonging to the same cluster.

**Keywords**—Clustering, Energy Market, Forecast, Linear Regression, Production Profile.

## INTRODUCTION

The Market Agent (MA) can trade the electricity produced by the assets through participation in energy markets. The day-ahead market allows the MA to submit buy and sell offers for the twenty-four hours of the following day, thus determining the prices and the traded energy. If the verified energy quantity, as indicated by the Transmission System Operator (TSO), differs from the scheduling in the Final Hourly Program (FHP) in the market, an excess or defect deviation occurs that may represent a payment obligation or or compensation for the Balance Responsible Party (BRP). Currently, imbalance settlement is carried out on an hourly basis. However, in March 2025, a change in the Imbalance Settlement Period (ISP) to quarter-hourly intervals is expected, which, according to some experts, may lead to an increase in imbalances [1]. Consequently, market agents must have access to the most accurate forecast possible, which is not always the case due to the mix of technologies represented in the market.

In this article, a clustering algorithm called K-Means is used to segment the assets of a MA based on predefined characteristics in the model, such as production, installed capacity, and geographical location. K-means is an unsupervised learning algorithm and one of the main analytical

methods in data mining. This algorithm classifies and segments data into clusters so that they are grouped according to similarity patterns. Although it is an easy to implement and fast method, it has some limitations in terms of computational complexity due to the need to calculate the distance between each object and the cluster centers in every iteration, reducing its efficiency.

The scientific community has been dedicated to analyzing this model. According to [2], the K-Means algorithm faces many challenges that negatively affect its performance in clustering. Since, during the initialization process, the user must define the number of clusters beforehand, this can influence the results and may lead to a suboptimal solution. To overcome this limitation, the article [3] presents a clustering algorithm, U-K-Means (Unsupervised K-Means), which addresses the initialization problem, as there is no need to identify the number of clusters in advance. U-K-Means determines the number of clusters, eliminating the need for other heuristic methods, such as the Elbow Method, to determine it. In [4], distributed computing mechanisms and clustering algorithms are used to extract features and classify load profiles based on their similarity. Neural network models were also employed within each cluster to test the forecasting model for various users in different regions. The results achieved show a high degree of accuracy, improving the forecasts. However, the use of temporal features from time series data is not explored. In [5], Machine Learning (ML) models and the k-Means algorithm are explored to obtain solar production forecasts for buildings. The authors concluded that using the clustering algorithm allows filtering large amounts of data, thereby speeding up the execution process of the models, making them more transparent and reliable. Solving forecasting problems requires processing, segmentation, and characterization of data samples to obtain a model with a high degree of accuracy.

Regarding forecasting models, these can be divided into two groups [6]: statistical methods and artificial intelligence methods. Statistical models are used when there is a set of data organized chronologically that shows the variation of a certain indicator over time, known as time series. The characteristics of the time series are related to the order and temporal dependency of their behavior. The series are ordered in time and are

---

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020

correlated with each other, meaning that data from previous periods can influence future data. In terms of behavior, the series is considered stationary if its statistical properties do not change over time. If there are variations, the series is considered non-stationary and may present trends or seasonal variations, such as in the case of production data, where there is a significant increase in energy produced during the summer months compared to the winter months. The most used statistical methods for generating forecast data include linear regression or multiple linear regression models, exponential models, autoregressive models (AR), Autoregressive Integrated Moving Average (ARIMA), among others. Artificial Intelligence (AI) methods are highly effective at detecting complex patterns that static methods cannot identify. The main methods used are Artificial Neural Networks (ANN), Support Vector Machines (SVM), deep learning, and Long Short-Term Memory (LSTM).

Given the research already developed in the application of clustering algorithms and forecasting methods, some limitations in the application of these models can be identified. Specifically, the analysis of time series characteristics combined with the identification of parameters that allow segmentation of production profiles based on other factors such as technology (hydroelectric, solar, wind), location, installed capacity, and type of installation (with or without self-consumption). Most of the existing literature is limited to analysing production profiles without establishing correlations with these indicators. This paper proposes the use of synthetic time-series generation models through the application of the K-Means algorithm, which allows extracting a set of similarity patterns from a finite group of production profiles [7], [8]. The application of this algorithm, together with the multiple linear regression model, improves forecasting accuracy by reducing deviations [9]. This methodology enables the segmentation of new installations within the AM portfolio based on existing clusters and the generation of time-series forecasts using historical data from installations in the same cluster. This approach makes the agent more autonomous in market bidding, as it eliminates the need to rely on an external forecasting provider.

## METHODOLOGY

This section presents the models applied in this article, and Fig. 1 shows the flowchart of the methodology applied. Initially, the input parameters were identified, the data was pre-processed, and the models identified in Fig. 1 were applied to generate the forecast for the target installation.

### A. Elbow Method

One of the limitations of the K-Means algorithm is determining the number of clusters. In [10], the Elbow Method is used to determine the ideal number of clusters. Initially, the Sum of Squared Errors (SSE) is calculated, and then the point where an inflection occurs. This value corresponds to the ideal number of clusters,  $k$ . The value of the Sum of Squared Errors (SSE) can be calculated using (1):

$$SSE = \sum_{k=1}^k \sum_{x_i \in S_k} \|x_i - C_k\|^2 \quad (1)$$

Where:

- $S_k$ , represents the set of points of the cluster,  $k$ .
- $x_i$ , it is a point of the cluster.
- $C_k$ , it is the centroid of the cluster,  $k$ .

The value of SSE decreases as  $k$  increases, because the points  $x_i$  are closer to the centroids  $C_k$ .

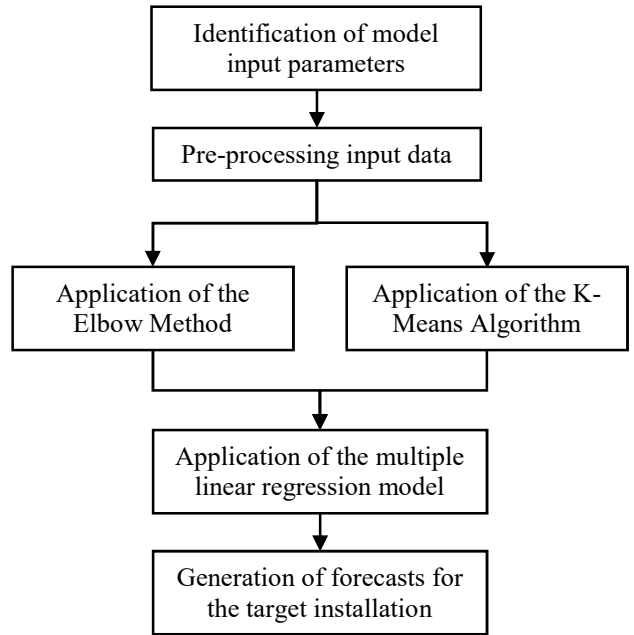


Figure 1. Flowchart of the models applied.

### B. K-Means

This algorithm starts with the definition of  $k$  and assigns the closest cluster to each point. Typically, this assignment is calculated using the Euclidean distance, which measures the distance between the points and the centroids of the clusters. The Euclidean distance can be obtained using (2),  $d(x_i, C_k)$ :

$$d(x_i, C_k) = \left[ \sum_{i=1}^n (x_i - C_k)^2 \right]^{\frac{1}{2}} \quad (2)$$

The K-Means algorithm can be executed multiple times due to the initialization of the number of clusters, especially if they are randomly selected. However, it does not guarantee that the most optimal solution to the problem will be found and may only provide a suboptimal result [11]-[14].

This algorithm can be executed as follows:

1. Selection of  $k$ , initial clusters, which represent the initial centroids.
2. Preliminary grouping, assigning the points from a given set and associating them with the closest centroid.
3. Recalculation of  $k$ , new centroids, based on the points previously assigned.
4. Steps 2 and 3 are repeated until the centroids no longer change.

### C. Linear Regression

The linear regression model [15],[16] is a technique used to analyze the relationship between a dependent variable and independent variables. If there is more than one independent variable, it is called multiple linear regression. This model uses a set of observations, training data, to predict the linear influence of these attributes on a target value. By applying the linear regression model, the data for the dependent variables are obtained, also known as test data.

The general equation of the multiple linear regression model (3):

$$y = a + b_1x_1 + \dots + b_nx_n + e \quad (3)$$

Where:

- $y$ , it is the dependent variable is the data we aim to predict, in the context of the problem, it will be the forecast.
- $x$ , it is the independent variable, in the context of the problem, it refers to the time series of production.
- $a$ , constant value, value of  $y$ , when  $x$  is 0.
- $b$ , it is the regression coefficient for each independent variable  $x$ .
- $e$ , it is the difference between the known value and the prediction.

### CASE STUDY

In the present case study, 65 geographically dispersed photovoltaic installations were considered, with full injection into the grid, meaning no self-consumption. The objective of this paper is to predict the production of a photovoltaic installation based on the actual production data of other installations within the same cluster by applying a clustering model and multiple linear regression. For the application of the models, Google Colab was used, allowing the execution of Python code. The first phase of the study involved defining the parameters to be analysed for the installations, such as installed capacity, ID code, and hourly real production data for each unit.

Based on these data, a database was created and imported into the code. The time series were pre-processed before implementing the models to ensure data reliability. The historical data period for each installation was analysed for the year 2024, followed by an assessment of missing data within the defined period. After this analysis, the presence of outliers, values significantly different from the rest, was examined. This step was crucial, as such data could reduce the accuracy of predictions. The previously mentioned indicators were then normalized. For a macro analysis of the data, the total production of all units throughout 2024 was summed, as shown in Fig. 2.

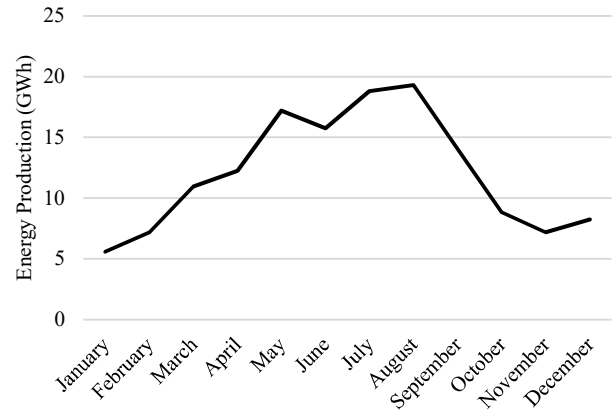


Figure 2. Total Energy Produced in 2024, in GWh.

Fig. 2, illustrates the seasonality inherent in the energy production of photovoltaic power plants, highlighting the trend in these time series. It is observed that during the winter months, the average energy production is around 7 GWh, whereas in the summer months, the average value rises to 17 GWh. In June, a decrease in energy production is observed compared to the previous month, which is related to the removal of an installation from the aggregator's portfolio, resulting in missing data.

Regarding the installed capacities, approximately 35 installations have 1 MW, 25 installations have capacities ranging between 1.2 and 1.5 MW, and 5 installations have capacities of 2.3, 8.6, 12.9, 17.8, and 18 MW. To geographically map the installations, the ID code indicator was used, which corresponds to the first four digits of the postal codes, with each ID corresponding to a municipality. This geographical breakdown is related to factors that directly influence the performance of the panels, namely ambient temperature and solar radiation. Fig. 3 presents the number of installations per ID code.

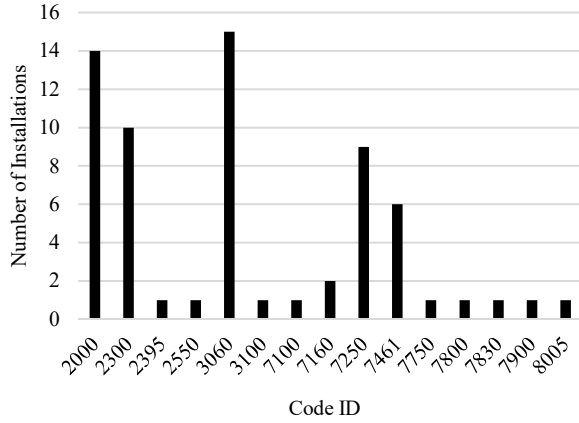


Figure 3. Number of Installations per ID Code.

### RESULTS

After preprocessing and importing the data, the methods described in Section II were applied. Initially, a data frame was created to initialize the variables under study (installed capacity, ID code, and the time series with production data in hourly format), followed by normalization.

The first algorithm applied was the Elbow Method to determine the optimal number of clusters. The algorithm was tested for different values of  $k$ , ranging from 1 to 10. For each  $k$  value, inertia was analysed, representing the proximity of points to the centroids. It was found that the optimal number of clusters is 6, as shown in Fig. 4.

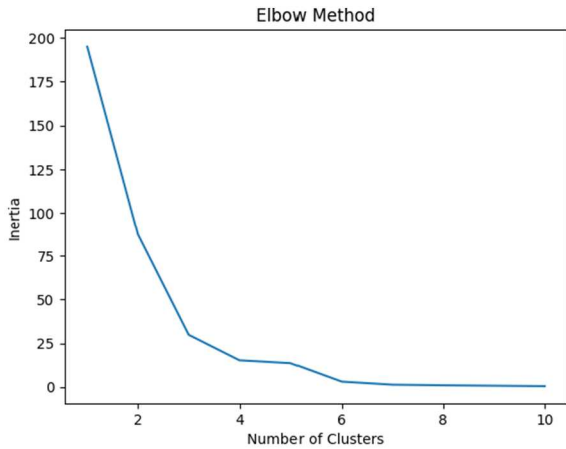


Figure 4. Application of the Elbow Method to Determine the Number of Clusters,  $k$ .

After determining the number of clusters, the K-Means method was applied to group the 65 installations into clusters. The input parameters used were the same as those in the Elbow Method. The algorithm assigned only one installation to clusters 3 and 5.

Although these installations have capacities of 18 MW and 17.8 MW, respectively, they are geographically distant from each other, with ID codes 3100 and 7150. In terms of representativeness, the clusters with the highest number of installations are clusters 1, 2, and 4, with 15, 20, and 26 installations, respectively. In these clusters, the average capacities are 1.1 and 1.2 MW, as shown in Table 1. In the case

of cluster 1, the ID code of the installations is 3060, in cluster 2, the ID codes range from 7250 to 7900, and in cluster 4, the codes range from 2000 to 2550. This breakdown demonstrates the correct distribution of installations across the clusters based on the input parameters. Based on the analysis of the production data, the target installation was determined, i.e., the installation for which the production profile will be generated through the application of the multiple linear regression model. The application of this model assumes the existence of independent variables, which, in the context of this problem, correspond to time series with historical production data.

TABLE I. Clusters Obtained through the Application of the K-Means Algorithm.

Number of Cluster	Number of Installations	Average Installed Capacity (MW)
0	2	10,8
1	15	1,2
2	20	1,1
3	1	18,0
4	26	1,1
5	1	17,8

The target installation belongs to cluster 1, as it is the installation with the smallest amount of historical data, and the purpose of this paper is to predict the energy produced. The goal is to use the existing data from the other 15 installations in this cluster to generate the energy production forecast. The historical real data of the target installation covers the period from 04/10/2024 to 24/11/2024, which will be used as training data. The goal is to predict the period from 25/11/2024 to 31/12/2024, which will serve as test data. The installations within this cluster with real data for the test period were analyzed, and only 4 installations contained data, as shown in Fig.5.

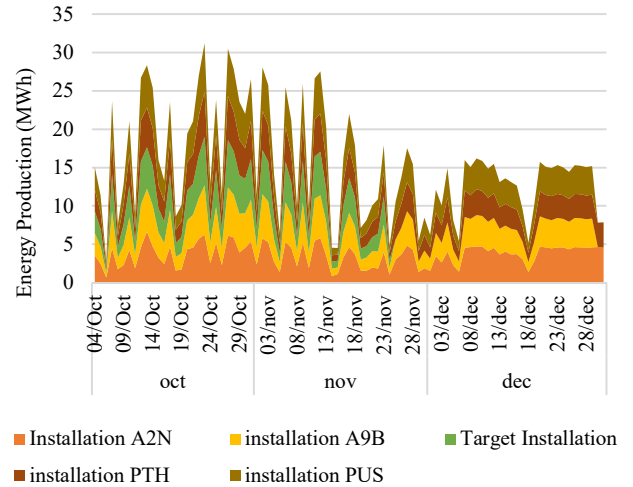


Figure 5. Production Data (MWh) of the Installations in Cluster 1.

Based on this data, the multiple linear regression model was applied, as there is more than one independent variable. The training and test data from the installations, along with the corresponding time periods, were defined. The predicted energy for the test period, obtained through the application of the model, was 112.57 MWh, while the actual production for the same period was 129.11 MWh. This indicates that the

overall prediction was overestimated, which corresponds to a receivable amount of €909.01 for the Aggregator (AM). The AM's forecast supplier predicted a production of 146.77 MWh, which translates to a payment obligation of €3,935.98.

The supplier's absolute deviation was 53.77 MWh, while with the application of the regression model, the deviation was 17.58 MWh, resulting in a 32% reduction. It is important to note that the regression model's effectiveness decreases after 28/12/2024, as the test data no longer includes the historical data of two installations due to missing data, as shown in Fig. 6.

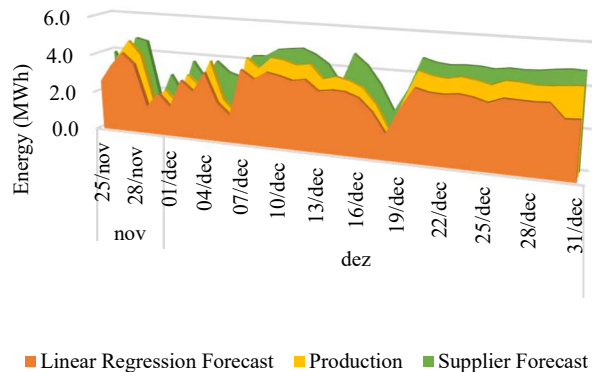


Figure 6. Test Data Obtained Through the Application of the Linear Regression Model.

A more detailed analysis of Fig. 5 reveals that the pattern of the regression model's forecast curve closely resembles the actual production. This is not the case with the supplier's forecast, particularly between December 4 and December 16, where a deviation of 13% from actual production is observed. This deviation can be reduced to 9% if a linear forecast is applied for the same period. Finally, some common indicators used in forecasting models were calculated, such as the Symmetric Mean Absolute Percentage Error (SMAPE), which was found to be 22%, and the Mean Absolute Error (MAE), which was 1.98%.

## CONCLUSIONS

This paper proposes the application of a clustering algorithm and a multiple linear regression model to predict the energy production of an installation. The application of this model resulted in a 32% reduction in absolute deviation compared to the forecast contracted by the Aggregator (AM) from a forecast provider. It is important to highlight that the greater the amount of training data, the more accurate the prediction will be. However, despite this limitation, the results obtained were highly positive in terms of deviation valuation, as the AM shifted from having to pay €3,935.98 to receiving €909.01. One of the main challenges of this approach was the limited amount of data available for training the model. Future work should incorporate historical data covering at least one year of production and apply Machine Learning methods to compare the effectiveness of statistical models with artificial intelligence techniques. This comparison would help assess the performance of each method under the same conditions.

## REFERENCES

- [1] T. Pinto, Z. Vale, "Electricity Market Participation Profiles Classification for Decision Support in Market Negotiation," in *Intelligent Data Mining and Analysis in Power and Energy Systems: Models and Applications for Smarter Efficient Power Systems*, IEEE, 2023, pp.171-186.
- [2] K. P. Sinaga and M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in *IEEE Access*, vol. 8, pp. 80716-80727, 2020.
- [3] A. Ikotun, A. Ezugwu, L. Abualigah, Belal Abuhaija, J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data", *Information Sciences*, Volume 622, 2023, pp 178-210.
- [4] H. Zhang, C. Zhu and S. Yang, "Multi-user Power Load Forecasting Based on K-means and Deep Neural Network," *2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI)*, Changchun, China, 2023, pp. 508-512.
- [5] B. Teixeira, L. Valina, T. Pinto, A. Reis, J. Barroso and Z. Vale, "Exploring Clustering to Improve Interpretability in Complex Energy Forecasting Models," *2024 International Conference on Smart Energy Systems and Technologies (SEST)*, Torino, Italy, 2024, pp. 1-6.
- [6] A. Setiawan, Z. Arifin, B. Sudiarto, F. H. Jufri, Q. Haramaini and I. Garniwa, "Comparison of Medium-Term Load Forecasting Methods (Splitted Linear Regression and Artificial Neural Networks) in Electricity Systems Located in Tropical Regions," *2022 3rd International Conference on Clean and Green Energy Engineering (CGEE)*, Istanbul, Turkey, 2022, pp. 84-88.
- [7] D. Viana, R. Teixeira, J. Baptista and T. Pinto, "Synthetic Data Generation Models for Time Series: A Literature Review," *2024 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, Sydney, Australia, 2024, pp. 1-6.
- [8] Y. Lu, Y. Zhang and S. Chen, "A Review of Time Series Data Mining Methods Based on Cluster Analysis," *2023 35th Chinese Control and Decision Conference (CCDC)*, Yichang, China, 2023, pp. 4198-4202.
- [9] W. Wang, G. Lyu, Y. Shi and X. Liang, "Time Series Clustering Based on Dynamic Time Warping," *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, China, 2018, pp. 487-490.
- [10] D. Marutho, S. Hendra Handaka, E. Wijaya and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," *2018 International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, 2018, pp. 533-538.
- [11] V. K. Dehariya, S. K. Shrivastava and R. C. Jain, "Clustering of Image Data Set Using K-Means and Fuzzy K-Means Algorithms," *2010 International Conference on Computational Intelligence and Communication Networks*, Bhopal, India, 2010, pp. 386-391.
- [12] S. R. L. B. M. J. P. K. S. Janani and R. R., "Fuzzy K-Means Clustering with ML Algorithm for Efficient Keyword Search," *2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL)*, Bhimdatta, Nepal, 2024, pp. 250-253.
- [13] S. Yu, Y. Yang and J. Zhao, "A Particle Swarm Optimization Model for Automatic Pricing and Replenishment of Vegetables using K-means Clustering Algorithm," *2024 Third International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE)*, Ballari, India, 2024, pp. 1-6.
- [14] R. Ban and Y. Deng, "Optimization Method for Predictive Models Based on ARIMA Time Series and K-means Clustering Algorithm," *2024 36th Chinese Control and Decision Conference (CCDC)*, Xi'an, China, 2024, pp. 2391-2396.
- [15] A. Setiawan, Z. Arifin, B. Sudiarto, F. H. Jufri, Q. Haramaini and I. Garniwa, "Comparison of Medium-Term Load Forecasting Methods (Splitted Linear Regression and Artificial Neural Networks) in Electricity Systems Located in Tropical Regions," *2022 3rd International Conference on Clean and Green Energy Engineering (CGEE)*, Istanbul, Turkey, 2022, pp. 84-88.
- [16] A. Jovanović, A. Krstić, S. Vujnović and Ž. Durović, "On Multivariate Linear Regression Applications," *2024 11th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN)*, Nis, Serbia, 2024, pp. 1-5.