Supervised Learning for Energy Forecasting in Power Systems

ERIC AKPOVIRORO OBAR, ABDELWAHED TOUATI, MAHMOUD ALMOSTAFA RABBAH, LAINCE PIERRE MOULEBE, NABILA RABBAH
Laboratory of Complex Cyber Physical Systems ENSAM,
Hassan II University,
Casablanca,
MOROCCO

Abstract: - Since the inception of the electric grid in the 19th century, power systems have continuously evolved due to technological, industrial, legislative, demographic, environmental, and economic factors. With the advent of machine learning, monitoring and anticipating the evolutionary trends of the electric grid has become possible. This is facilitated by the convergence of vast data availability, sophisticated algorithms, and advanced computational capabilities. Our focus is on utilizing the supervised learning paradigm of machine learning for predictive analytics in power systems. Specifically, we aim to forecast electricity consumption, leveraging the predictive power of supervised learning techniques.

Key-Words: - Machine learning, Linear regression, Supervised learning, Trends, Exploratory Data Analysis (EDA), Homoscedasticity, R-squared, Forecasting.

Received: May 27, 2024. Revised: November 29, 2024. Accepted: December 23, 2024. Published: March 28, 2025.

1 Introduction

Supervised learning is a branch of machine learning that involves training a model on a labeled dataset, where the correct output is provided for each example in the training set. The goal of supervised learning is to build a model that can make predictions based on new, unseen examples.

In the power sector, supervised learning can be used for energy forecasting by training a model to predict future electricity demand or generation based on historical data. To achieve this, one would need to first gather a dataset of historical electricity demand or generation data, along with any relevant features or variables that might influence demand or generation.

With the evolution of the electric grid due to technological, industrial, legislative, demographic environmental, and economic factors, the predictive analytics of power systems is of vital importance for the strategic planning and balance of electricity demand and supply. Thanks to our era and digital age which guarantees the availability of large amounts of data and the high computational capacity of computers, statistical models and algorithms can be trained to make predictions or classifications in sectors of our choosing. Our interest in this paper would be the application of the supervised learning technique for the predictive analytics of electricity consumption in Cote d'Ivoire (Figure 1).

Fig. 1: Map of Côte d'Ivoire, [1]

1.1 Literature Review of Some Supervised Learning Techniques

In recent years, accurately predicting energy demand has become more important than ever, especially with the increasing reliance on renewable energy sources within the power grid, [2], [3]. To tackle this challenge, researchers have explored various supervised learning methods. These methods include well-known techniques like linear regression, random forest, gradient boosting, support vector regression (SVR), k-nearest neighbors (KNN), and decision trees.

Linear regression is one of the first methods that comes to mind when dealing with supervised learning techniques. This method maps input and output data based on a dataset of input and output pairs, [4]. It is a classic and effective approach especially for long-term energy forecasting, [5]. A more robust technique is the Random forest method. It does combine several decision trees to ensure the reliability and accuracy of predictions, [2], [3].

Metrics like the Mean Absolute Error(MAE) amongst others, show that Random forest models outperform some machine learning techniques like the Auto Regressive Moving Average Models (ARMA) presenting minimal prediction errors in energy forecasting applications, [6]. We also have the Gradient Boosting ensemble method which combines weak models of decision trees and iteratively corrects errors of previous decision trees through training. This ultimately results in a more robust predictive model, [7].

Next is the Support Vector Regression (SVR) which stems from the Support Vector Machines (SVM). It is a kernel based method used for short-term energy forecasting capable of handling complex and non-linear relationships, [7]. Decision trees are a popular machine learning algorithm that recursively partitions the inputspace to make predictions, and has been widely used for energy forecasting due to their interpretability and flexibility, [2], [3], [4], [7], [8]. K-nearest neighbors is a memory-based algorithm that relies on the similarity of historical data to make predictions and has been applied to energy forecasting with promising results, [4], [9].

Recent research has also explored the use of deep learning techniques for energy forecasting, which have shown promising results in capturing the inherent nonlinear and complex patterns in renewable energy data,[2]. However, the effectiveness of these supervised learning techniques can vary depending on the specific characteristics of the energy data, the forecasting horizon, and the application requirements.

Our goal is to guarantee strategic planning and balance of electricity demand and supply by accurately predicting/forecasting electricity consumption. And Figure 2 depicts our research methodology. We began by collecting and cleaning a sufficient amount of historical data (Primary Data gotten from **CI-Energies** beginning from1985-2021 on electricity production and consumption in Cote d’Ivoire (Precisely, the regional directorate of Yopougon). After which, we Split the data into training and test sets. We then used different supervised learning methods to train our model so as to capture the underlying patterns and trends in our data. Then, we evaluated the model performance using the test set and made the necessary adjustments in order to fine-tune the model. We validated our model using additional data. This allowed us to ensure that our model provided reliable predictions.

2 Research Methodology

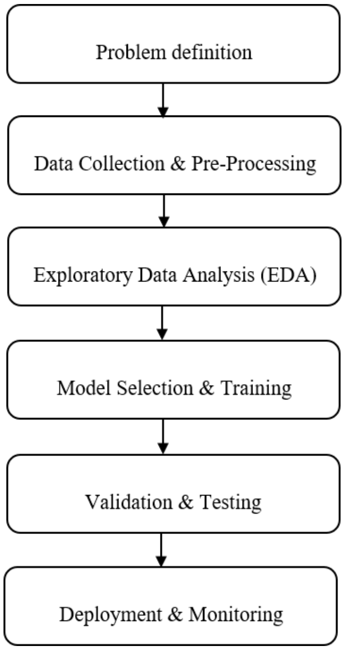

Problem definition
Data Collection & Pre-Processing
Exploratory Data Analysis (EDA)
Model Selection & Training
Validation & Testing
Deployment & Monitoring

Fig. 2: Research Methodology

3 Results

3.1 Exploratory Data Analysis (EDA)

Our dataset spans from the year 1985 to 2021 with information on monthly energy consumption in MWh. And no missing values were observed in our dataset. The columns that represent key information in our dataset are: Year Index, Regional Key, monthly energy consumption (MWh), and the date.

Thanks to our EDA, we notice the exponential increase of electricity consumption as shown in Figure 3, Figure 4 and Figure 5. Figure 6 illustrates the monthly energy distribution and Figure 7 presents the correlation matrix and Table 1 presents statistical summary of energy consumption of the regional directorate of Yopougon.

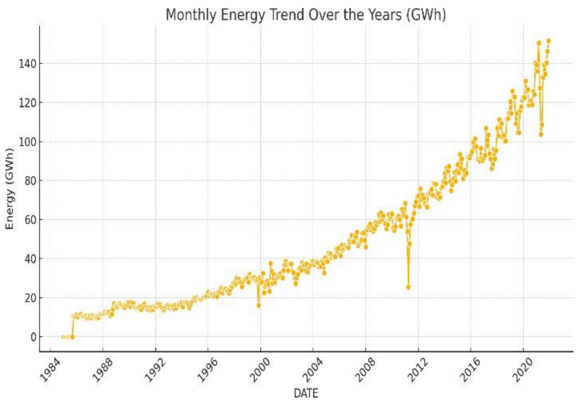


Fig. 3: Monthly Energy Trend Over the Years (GWh)

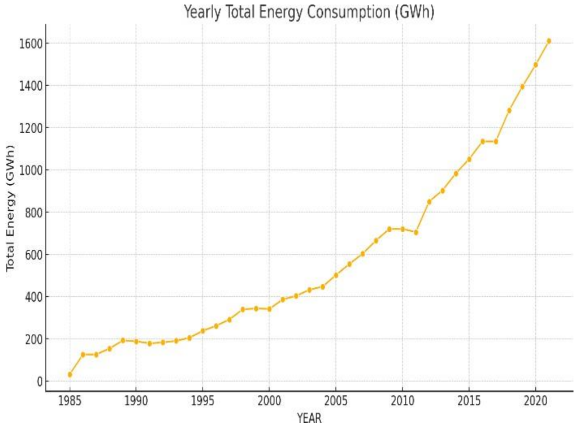


Fig. 4: Yearly Total Energy Consumption (GWh)

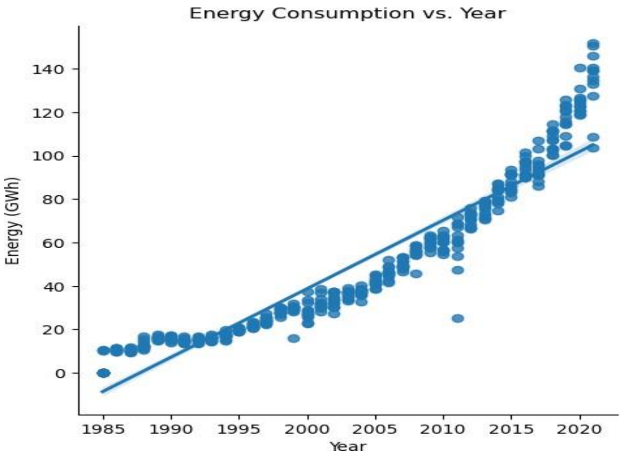


Fig. 5: Energy Consumption Vs. Year   Fi

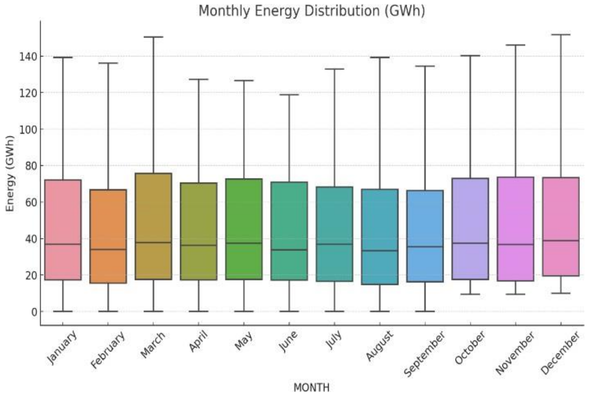


Fig. 6: Monthly Energy Distribution (GWh)

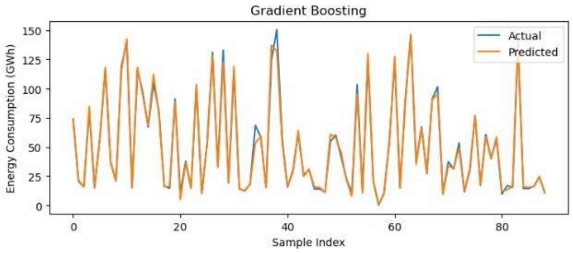


Fig. 7: Gradient Boosting

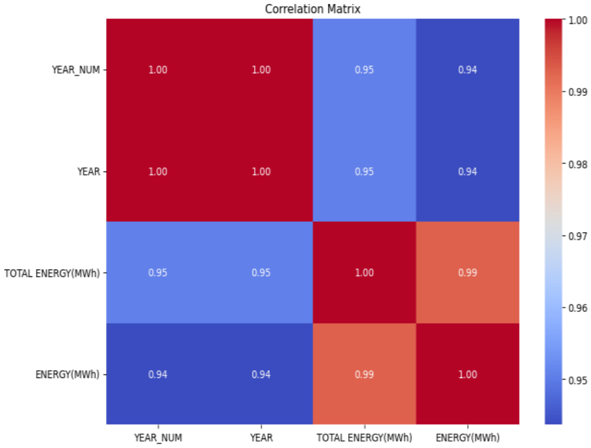


Fig. 8: Correlation Matrix

Table 1. Statistics Summary

Statistics	Energy (GWh)
Mean	48.11
Std	35.74
Min	00.00
25%	17.33
50%	36.82
75%	71.83
Max	151.16

3.2 Model Selection & Training

Our model selection was predicated on the exploration of several supervised learning techniques with the goal of having the most reliable model for energy forecasting in the regional directorate of Yopougon in Côte d'Ivoire. The models explored were Linear Regression, Random Forest, Gradient Boosting, Support Vector Regression, k-nearest Neighbors (KNN), and Decision Tree Regression models. These models were assessed with metrics such as the Mean Absolute Error(MAE), the Mean Squared Error(MSE), the Root Mean Squared Error(RMSE), and the R-squared metrics. Graphical representations of the various models mentioned above are given below.

Our findings did reveal the Gradient Boosting model as the best-performing model delivering a result with the smallest MAE and RMSE of 2.28 and 3.92 respectively.

The next best-performing model according to our analysis is the Random Forest model with MAE and RMSE of 2.48 and 4.13 respectively.

The Random Forest model is followed by Decision Tree Regression with MAE and RMSE of 2.83 and 5.32 respectively.

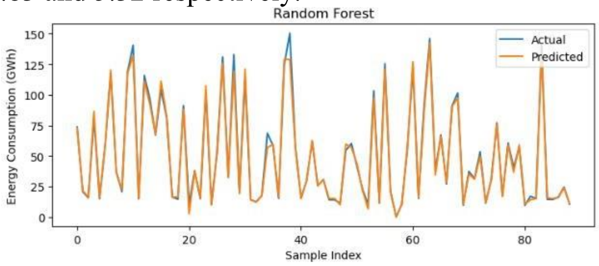


Fig. 9: Random Forest

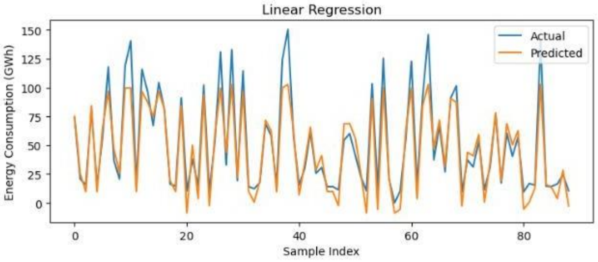


Fig. 10: Linear Regression

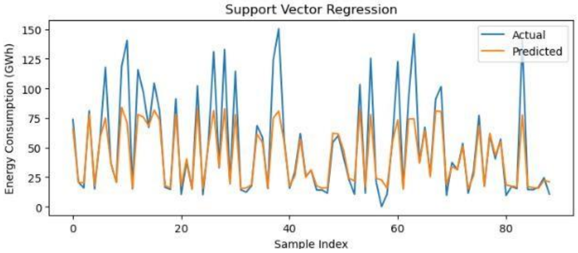


Fig. 11: Support Vector Regression

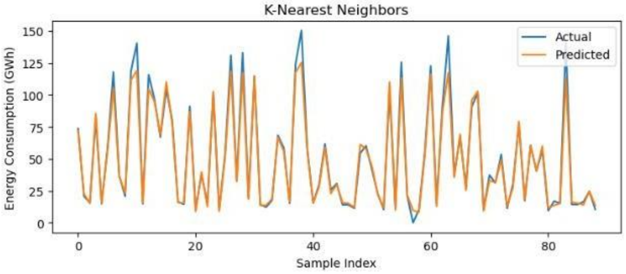


Fig. 12: K-Nearest Neighbors

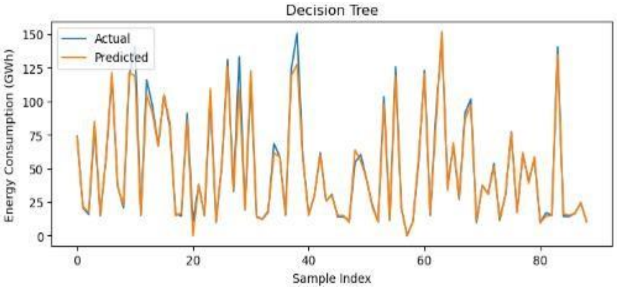


Fig. 13: Decision Tree

Table 2. Performance metrics of each model

Model	**MAE**	**MSE**	**RMSE**	**R-squared**
Linear Regression	10.75	206.90	14.38	0.883014
Random Forest	2.48	17.03	4.13	0.990369
Gradient Boosting	2.28	15.36	3.92	0.991315
SVR	11.73	462.53	21.51	0.738472
KNN	3.85	46.85	6.84	0.97351
Decision Tree	2.83	28.35	5.32	0.98397

4 Discussion (Model Validation & Testing)

Table 2 and Figure 8, Figure 9, Figure 10, Figure 11, Figure 12 and Figure 13 do give the results of the different models tested using metrics such as the MAE, MSE, RMSE, and R-squared values. In the sections below we will discuss the performance evaluation, assumptions, and diagnostics of the different models.

4.1 Performance Evaluation

1) Linear Rrogression:

The Linear Regression Model did produce an R-squared value of 0.883. This means that our model was able to explain about 88.3% of the variation in the energy consumption data. However, the model exhibited a relatively high RMSE of 14.38, suggesting that the predictions were not as precise as those of other models. While linear regression provides a straightforward approach, its performance was suboptimal compared to more advanced techniques.

F-test: The F-statistic for the model was 1446 with a p-value of 1.74e-170, indicating that our model is highly significant and that our explanatory variable (Year) is significantly related to the dependent variable(Energy).

2) The Random Forest:

The Random Forest model demonstrated excellent predictive capabilities, with an R- squared value of 0.990 and a low RMSE of

4.13. This model effectively captured the complex patterns in the data, resulting in highly accurate predictions. The robustness of the Random Forest algorithm, which aggregates multiple decision trees, likely contributed to its superior performance. Our findings are confirmed, [10], [11], [12].

3) Gradient Boosting:

Gradient Boosting emerged as the best-performing model, achieving the highest R- R-squared value of 0.991 and the lowest RMSE of 3.92. This model's ability to iteratively improve its predictions by focusing on errors from previous iterations allowed it to outperform other models. The slight edge over Random Forest underscores the effectiveness of boosting techniques in handling non-linear relationships and interactions within the data. Its attributes make for forecasting accuracy for both short-term and long-term energy predictions, [13].

4) Support Vector Regression (SVR):

SVR showed the weakest performance among the evaluated models, with an R-squared value of 0.738 and a high RMSE of

21.51. The SVR model's inability to capture the underlying patterns in the data suggests that it may not be well-suited for our dataset, possibly due to its sensitivity to the choice of kernel and hyperparameters.

5) K-Nearest Neighbours (KNN):

The KNN model achieved a respectable R-squared value of 0.974 and an RMSE of 6.84. While not as accurate as Random Forest or Gradient Boosting, KNN still provided reasonably good predictions. Its performance demonstrates the potential of instance-based learning methods in regression tasks, though it is slightly less effective in capturing complex patterns compared to ensemble methods. While k-nearest neighbors may not outperform more complex machine learning algorithms, it can be a useful tool in hybrid forecasting models, where it is combined with linear statistical models to leverage the strengths of both approaches.

6) Decision Tree:

The Decision Tree model performed well, with an R-squared value of 0.984 and an RMSE of 5.32. Although it did not match the accuracy of Random Forest or Gradient Boosting, it still outperformed Linear Regression and SVR. Decision Trees are known for their interpretability and ability to model non-linear relationships, which likely contributed to their solid performance. Our results from the decision tree model do show that they are suitable choices for energy forecasting applications, [14].

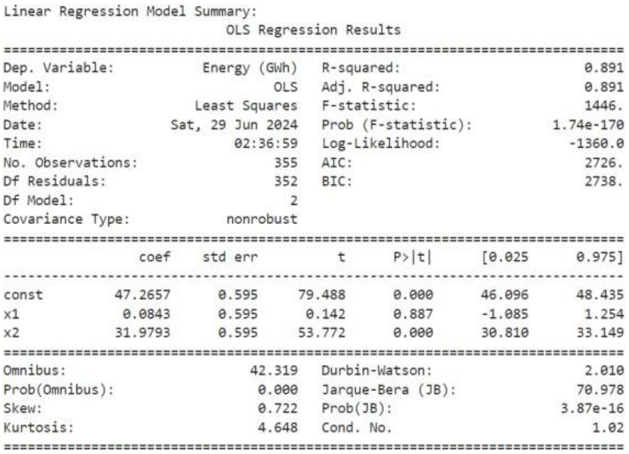
7) Assumptions and Model Diagnostics:

For the Linear Regression model, diagnostic checks were conducted to ensure the assumptions of linear regression were met. The residual plots indicated no significant patterns, suggesting that the assumptions of linearity and homoscedasticity were reasonably satisfied. The Q-Q plot confirmed that the residuals were approximately normally distributed, and the Durbin-Watson statistic (1.76) suggested no significant autocorrelation in the residuals. Additionally, the F-test (F-statistic: 1446, p-value: 1.74e- 170) indicated that the overall model was highly significant.

- **Homoscedasticity**: The residual plot for the Linear Regression model showed that the residuals were randomly scattered around the horizontal axis, without forming a discernible pattern. This suggests that the assumption of homoscedasticity (constant variance of residuals) is reasonably satisfied. Homoscedasticity is essential for ensuring that the model's predictions are reliable across all levels of the independent variables.

However, it is important to note that the F-test and certain assumptions, such as homoscedasticity, are specific to linear regression and do not directly apply to non-linear models like Random Forest, Gradient Boosting, SVR, KNN, and Decision Tree. For these models, performance metrics such as MAE, MSE, RMSE, and R-squared provide a comprehensive evaluation of their predictive capabilities and Figure 14 provides a summary of the linear regression model

In conclusion, the Gradient Boosting and Random Forest models demonstrated the best performance in predicting monthly energy consumption, with Gradient Boosting having a slight edge. These models effectively captured the complex patterns in the data, resulting in highly accurate predictions. The findings highlight the importance of selecting appropriate models for specific datasets, as advanced ensemble methods significantly outperformed traditional linear regression and other non-linear approaches. Future work could explore hyperparameter optimization and feature engineering to further enhance model performance. The significance of the F-test in the linear regression model allows us to confidently reject the null hypothesis, reinforcing the model's validity and reliability in capturing the relationship between the predictors and energy consumption.



```
Linear Regression Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:            Energy (GWh)   R-squared:                      0.891
Model:                             OLS   Adj. R-squared:                 0.891
Method:                  Least Squares   F-statistic:                    1446.
Date:                 Sat, 29 Jun 2024   Prob (F-statistic):         1.74e-170
Time:                         02:36:59   Log-Likelihood:               -1360.0
No. Observations:                  355   AIC:                            2726.
Df Residuals:                      352   BIC:                            2738.
Df Model:                            2
Covariance Type:             nonrobust
==============================================================================
                coef    std err          t      P>|t|      [0.025       0.975]
------------------------------------------------------------------------------
const        47.2657      0.595     79.488      0.000      46.096       48.435
x1            0.0843      0.595      0.142      0.887      -1.085        1.254
x2           31.9793      0.595     53.772      0.000      30.810       33.149
==============================================================================
Omnibus:                        42.319   Durbin-Watson:                  2.010
Prob(Omnibus):                   0.000   Jarque-Bera (JB):              70.978
Skew:                            0.722   Prob(JB):                    3.87e-16
Kurtosis:                        4.648   Cond. No.                        1.02
==============================================================================
```

Fig. 14: Linear Regression Model Summary

Acknowledgement:

We wish to express our sincere appreciation to CI-Energies, led by its distinguished Director General, Mr. Noumory SIDIBE, for the invaluable contribution of providing the energy data that greatly supported our research in Côte d'Ivoire.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the authors used Chatgpt 4 in order to enhance the readability of the manuscript. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication

References:

[1] "Cote d'Ivoire Maps & Facts - World Atlas", [Online]. https://www.worldatlas.com/maps/cote-d-ivoire (Accessed Date: Dec. 17, 2024).

[2] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Convers. Manag.*, vol. 198, no. April, p. 111799, 2019, doi: 10.1016/j.enconman.2019.111799.

[3] A. Mellit, A. M. Pavan, E. Ogliari, S. Leva, and V. Lughi, "Advanced methods for photovoltaic output power forecasting: A review," *Appl. Sci.*, vol. 10, no. 2, pp. 1–23, 2020, doi: 10.3390/app10020487.

[4] T. Hong, P. Pinson, Y. Wang, R. Weron, D. Yang, and H. Zareipour, "Energy Forecasting: A Review and Outlook," *IEEE Open Access J. Power Energy*, vol. 7, no. Xx, pp. 376–388, 2020, doi: 10.1109/OAJPE.2020.3029979.

[5] M. A. Hammad, B. Jereb, B. Rosi, and D. Dragan, "Methods and Models for Electric Load Forecasting: A Comprehensive Review," *Logist. Sustain. Transp.*, vol. 11, no. 1, pp. 51–76, 2020, doi: 10.2478/jlst-2020-0004.

[6] Y. Lu ., "The application of improved random forest algorithm on the prediction of electric vehicle charging load," *Energies*, vol. 11, no. 11, 3207, 2018, doi: 10.3390/en11113207.

[7] N. G. Paterakis, E. Mocanu, M. Gibescu, B. Stappers, and W. Van Alst, "Deep learning versus traditional machine learning methods for aggregated energy demand prediction," *2017 IEEE PES Innov. Smart Grid Technol. Conf. Eur. ISGT-Europe 2017 - Proceedings, Turin, Italy*, vol. 2018-Janua, pp. 1–6, 2017, doi: 10.1109/ISGTEurope.2017.8260289.

[8] P. A. Schirmer, I. Mporas, and I. Potamitis, "Evaluation of Regression Algorithms in Residential Energy Consumption Prediction," *Proc. - 2019 3rd Eur. Conf. Electr. Eng. Comput. Sci. EECS 2019*, Athens, Greece pp. 22–25, 2019, doi: 10.1109/EECS49779.2019.00018.

[9] T. Ahmad and H. Zhang, "Novel deep supervised ML models with feature selection approach for large-scale utilities and buildings short and medium-term load requirement forecasts," *Energy*, vol. 209, p. 118477, 2020, doi: 10.1016/j.energy.2020.118477.

[10] H. Alharkan, S. Habib, and M. Islam, "Solar Power Prediction Using Dual Stream CNN-LSTM Architecture," *Sensors*, vol. 23, no. 2, pp. 1–12, 2023, doi: 10.3390/s23020945.

[11] R. Bonetto and M. Rossi, 2017, "Machine Learning Approaches to Energy Consumption

Forecasting in Households," 10.48550/arXiv.1706.09648. pp. 6–9.

[12] M. Negnevitsky, P. Mandal, and A. K. Srivastava, "Machine learning applications for load, price and wind power prediction in power systems," *2009 15th Int. Conf. Intell. Syst. Appl. to Power Syst. ISAP '09*, 2009, Curitiba, Brazil, doi: 10.1109/ISAP.2009.5352820.

[13] Z. Guo, K. Zhou, X. Zhang, and S. Yang, "A deep learning model for short-term power load and probability density forecasting," *Energy*, vol. 160, pp. 1186–1200, 2018, doi: 10.1016/j.energy.2018.07.090.

[14] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably Unequal? The Effects of Machine Learning on Credit Markets," *J. Finance*, vol. 77, no. 1, pp. 5–47, 2022, doi: 10.1111/jofi.13090.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No sources of funding

Conflict of Interest

The authors have no conflicts of interest to declare.