Journal of Artificial Intelligence and Capsule Networks (ISSN: 2582-2012)
www.irojournals.com/aicn/

Sustainable Energy Transition: Analyzing the Impact of Renewable Energy Sources on Global Power Generation

Rahul Kumar Jha

Department of Electrical Engineering, Pashchimanchal Campus, Tribhuvan University, Pokhara, Nepal

Email: rahul.752418@pasc.tu.edu.np

Abstract

This study delves into the intricate relationship between power plant attributes and electricity generation, employing data analysis and predictive modelling techniques. Through a comprehensive analysis of a global power plant dataset, critical factors such as plant capacity and commissioning year were identified as significant influencers on electricity generation. The research utilized correlation heatmaps to visually represent these relationships, offering valuable insights for policymakers and investors. A linear regression model was employed, leveraging capacity and commissioning year as features to predict electricity generation. The model's accuracy was evaluated using mean squared error, providing a quantitative measure of its predictive capabilities.

Keywords: Regression, Machine Learning, Predictive Analysis, Correlation Heatmap, Global Energy Data, Data Cleaning

1. Introduction

With a greater focus on renewable energy sources, the world's energy landscape has recently undergone a radical transformation toward sustainability. The investigation and implementation of renewable energy technologies have become of utmost significance as the world struggles with the problems caused by climate change and the need to cut carbon emissions. This paradigm change toward environmentally friendly energy sources not only

DOI: https://doi.org/10.36548/jaicn.2023.3.007
Received: 02.09.2023, received in revised form: 21.09.2023, accepted: 05.10.2023, published: 19.10.2023

responds to environmental issues, but also holds the key to guaranteeing energy security and promoting economic growth[1].

This research effort is placed at the nexus of technical innovation and environmental stewardship in a world where the demand for clean, dependable, and sustainable energy solutions is growing[2]. This study seeks to make a significant contribution to the ongoing conversation about creating a more sustainable energy future for future generations by objectively assessing the effect of renewable energy sources on global power generation.

1.1 Background

The traditional fossil fuel-based energy systems are being replaced by renewable energy sources, which has created a fundamental shift in the global energy landscape. The urgent need to combat climate change, lower greenhouse gas emissions, and assure a future powered by sustainable energy sources is what is driving this transition. This shift is taking place against a backdrop of increased environmental concerns, fluctuating fossil fuel costs, and rising energy consumption, particularly in emerging economies. The efficiency and affordability of renewable energy technologies, such as solar photovoltaics, wind turbines, hydropower plants, and geothermal systems, has significantly increased[3].

As a result, renewable energy now play a crucial role in both national and global energy policies. Aiming to reduce carbon emissions, initiatives like the Paris Agreement have compelled countries to invest in infrastructure for renewable energy sources and investigate novel ways to incorporate renewables into current energy systems[4], [5]. In addition, incorporating renewable energy sources into electricity generation has broad ramifications. It not only diversifies the energy mix but also enhances energy security by reducing dependence on finite fossil fuels.

The promotion of a green economy, the creation of job opportunities, and the encouragement of research and development in the field of renewable energy all contribute to economic growth. Despite these developments, problems still exist. Widespread use of renewable energy is hampered by problems with grid integration, energy storage, regulatory frameworks, and public perception. Therefore, it is crucial to have a thorough grasp of how renewable energy sources affect the world's power production. These insights are essential for helping decision-makers, energy stakeholders, and environmentalists create successful plans, fix problems, and quicken the speed of the transition to sustainable energy[6].

1.2 Research Objectives

- **Data Exploration and Cleaning**

1. Investigate the dataset's structure and identify missing values.
2. Remove incomplete rows to ensure data reliability and completeness.

- **Correlation Analysis**

1. Explore correlations between key variables (e.g., capacity, commissioning year, electricity generation) using correlation heatmaps.
2. Understand the relationships between these variables to gain insights into their interdependencies.

- **Analyze the practical implications of the model's findings for policymakers and investors in the energy sector.**

2. Methodology[7]–[11]

1. Data Preprocessing

(i) Exploratory Data Analysis (EDA): A comprehensive analysis of the dataset's features and distributions was conducted to gain insights into the data's nature and structure. Visualization techniques were employed to understand the inherent patterns within the dataset.

(ii)Data Cleaning: Incomplete or missing data points, especially in crucial columns like 'capacity_mw', 'commissioning_year', and 'generation_gwh_2017', were identified and subsequently removed. This step ensured the dataset's cleanliness and eliminated potential biases.

2. Correlation Analysis

(i) Selection of Numeric Variables: Relevant numeric variables, namely 'capacity_mw', 'commissioning_year', and 'generation_gwh_2017', were chosen for correlation analysis. These variables were considered significant due to their potential influence on electricity generation.

(ii) Correlation Heatmaps: Utilizing the seaborn library, correlation heatmaps were generated. These visual representations enabled a clear understanding of the relationships between the selected variables. Positive or negative correlations between variables were explored, providing valuable insights into potential correlations and their strengths.

3. Predictive Modelling

- **Feature Selection:** 'Capacity_mw' and 'commissioning_year' were selected as predictor variables based on their relevance to electricity generation. 'Generation_gwh_2017' was designated as the target variable, representing the outcome of interest.

- **Data Splitting:** The dataset was divided into training (80%) and testing (20%) sets using the train_test_split function from scikit-learn. This ensured an adequate amount of data for model training while retaining a subset for evaluation purposes.

- **Linear Regression Model:** A linear regression model, implemented using the scikit-learn library, was trained on the training dataset. This model aimed to predict electricity generation in 2017 based on the selected features. The linear regression algorithm was chosen due to its simplicity and interpretability.

- **Model Evaluation:** Mean Squared Error (MSE) was employed as the evaluation metric. MSE quantifies the average squared differences between the actual and predicted values, providing a measure of the model's accuracy. A lower MSE indicated a better fit of the model to the data.

By following this systematic methodology, the analysis ensured a rigorous examination of the dataset, thoughtful feature selection, and an appropriate evaluation technique, laying the foundation for robust insights into the factors influencing electricity generation and the performance of the predictive model.

3. Nature of Dataset[12]

There are 34,936 individual samples in the dataset. Each data point represents a specific instance or observation collected for analysis.

The nature of the dataset used in the given Python program appears to be related to global power plants. While the exact details of the dataset's content are not provided in the code, the analysis implies the presence of several key columns:

1. **Capacity (capacity_mw):** This likely represents the power generation capacity of each power plant, measured in megawatts (MW). It indicates the maximum amount of electricity the plant can produce under optimal conditions.

2. **Commissioning Year (commissioning_year):** This column likely contains the year when each power plant became operational. Knowing the commissioning year provides insights into the age of the plant, which can be a crucial factor in determining its efficiency and generation output.

3. **Electricity Generation in 2017 (generation_gwh_2017):** This column likely denotes the actual electricity generation of each power plant in gigawatt-hours (GWh) during the year 2017. It represents the real-world output of the power plants for that specific year.

Given the presence of these columns, the dataset appears to be related to the energy sector, specifically focusing on power plant capacities, commissioning years, and electricity generation figures. This type of dataset is common in studies related to energy production, efficiency analysis, and predictive modeling for electricity generation trends.

3.1 Type of Linear Regression

- **Multiple Linear Regression:** Multiple linear regression is employed because there are multiple independent variables (features) involved in predicting electricity generation. Each feature contributes to the prediction of the target variable.

- **Multiple Linear Regression Model:** In multiple linear regression, the relationship between the dependent variable (y) and n independent variables $(x_1,x_2,...,x_n)$ is expressed as a linear equation:

  $$y=b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_n \cdot x_n + \varepsilon$$

  Where:

y is the dependent variable (electricity generation in GWh).

x_1, x_2,,x_n are the independent variables (features such as capacity, commissioning year, etc.).

b0 is the intercept (constant term).

b_1,b_2,...,b_n are the coefficients (weights) of the independent variables, indicating their impact on y.

ε represents the error term, accounting for unexplained variance and randomness in the data.

The linear regression model estimates the coefficients (b_0,b_1,...,b_n) during the training process to minimize the difference between the predicted ($\hat{y}$) and actual (y) values. The objective is to find the best-fitting line that minimizes the sum of squared errors (residuals) between predicted and actual values.

In the provided Python code, the '**LinearRegression**' class from scikit-learn library is used, which internally implements the multiple linear regression algorithm. The fit() function of the '**LinearRegression**' class computes the coefficients (b_0,b_1,...,b_n) based on the provided training data.

In the provided code, regularization techniques like Lasso (L1 regularization) and Ridge (L2 regularization) are not explicitly used. The code focuses on handling missing values, performing feature selection, and training a Gradient Boosting Regressor model without applying regularization.

3.2 Proposed Algorithm and Flowchart

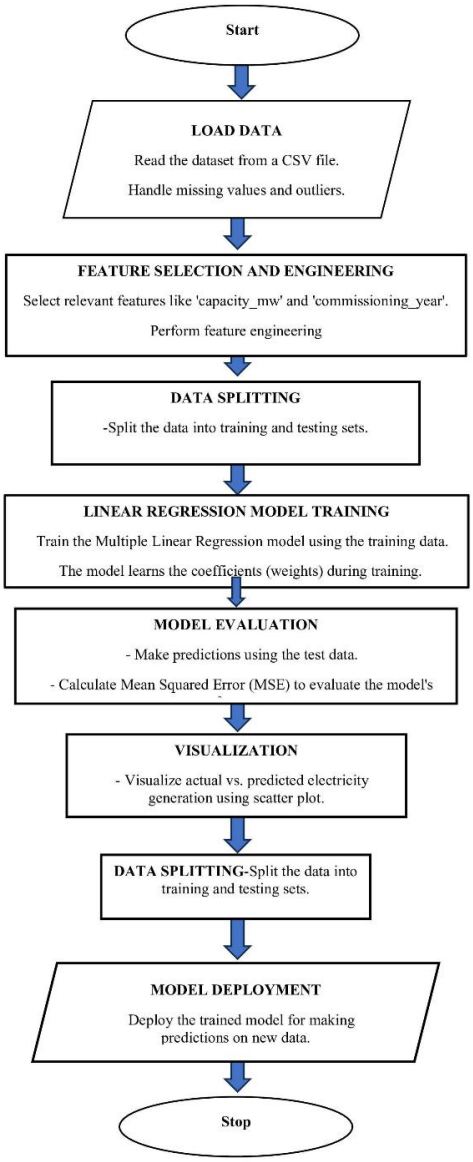

Figure 1. Proposed Flowchart

Libraries used in the Program[5], [13], [14]

Table 1. Libraries Used

Library	Responsibilities
pandas	Responsible for data manipulation and analysis. It provides data structures like DataFrame to work with structured data, allowing cleaning, filtering, and transforming data.
Matplotlib.pyplot	Provides a comprehensive set of plotting functions, enabling the creation of static, interactive, and animated visualizations in Python. In this context, it's used for creating various plots, including correlation heatmaps and scatter plots.
Seaborn	Built on top of Matplotlib, Seaborn is used for statistical data visualization. It simplifies the process of creating aesthetically appealing visualizations, such as heatmaps, which are crucial for understanding correlations within the dataset.
Sklearn.model_selection.train_test_split	A function from the scikit-learn library used for splitting datasets into train and test sets. It enables the evaluation of machine learning models by training them on a subset of the data and testing their performance on another subset.
Sklearn.linear_model.LinearRegression	Part of scikit-learn, it provides an implementation of linear regression, a fundamental machine learning algorithm used for modelling the relationship between a dependent variable (in this case, electricity generation) and one or more independent variables (features like capacity and commissioning year).
Sklearn.metrics.mean_squared_error	Used for calculating the mean squared error, which measures the average squared differences between the actual and predicted values. It's a crucial metric for evaluating the performance of regression models, providing insights into prediction accuracy.

3.3 Modifications to Meet Research Requirements

1. Feature Selection: Relevant features such as ‘capacity_mw’ (capacity of the power plant) and ‘commissioning_year’ (year the plant was commissioned) are selected based on their potential impact on electricity generation. These features are considered as it is directly related to the generation capacity and the age of the power plant, which can influence the electricity output.

2. Feature Engineering: The code calculates the ‘plant_age’ by subtracting the ‘commissioning_year’ from the current year (2023). This derived feature captures the age of the power plant, which can be crucial in predicting electricity generation. Including engineered features can enhance the model’s predictive capability.

3. Data Preprocessing: Missing values in ‘capacity_mw’ and ‘commissioning_year’ are handled using imputation. Outliers in the ‘generation_gwh_2017’ feature are removed to ensure that extreme values do not disproportionately influence the regression line.

4. Model Evaluation: Mean Squared Error (MSE) is employed as the evaluation metric. MSE quantifies the average squared difference between predicted and actual electricity generation. Lower MSE values indicate better model performance.

5. Visualization: The program includes a scatter plot to visualize the actual vs. predicted electricity generation. Visualization aids in understanding the model’s performance by comparing predicted values with the actual data points.

4. Result

The correlation heatmap generated in this analysis plays a pivotal role in understanding the relationships between different variables within the dataset. A correlation heatmap is a graphical representation of the correlation matrix, displaying the strength and direction of relationships between variables. In this specific case, the heatmap was used to visualize correlations between numerical variables such as power plant capacity, commissioning year, and electricity generation in 2017.

```
  data = pd.read_csv('C:\\Users\\Rahul\\Desktop\\global_power_plant_database.csv')
Mean Squared Error: 1145390.2796527164
R-squared: 0.20309912275529474
```

Figure 2. MSE of the Predicted Model

The mean squared error (MSE) of the model's predictions, which is approximately 1,145,390.28, and the R^2 score, which is approximately 0.203.

The heatmap offers several advantages in the analysis process:

1. Identifying Patterns: By visualizing correlations, analysts can quickly identify patterns in the data. Positive correlations (values close to 1) indicate variables that increase or decrease together, while negative correlations (values close to -1) suggest variables that move in opposite directions.

2. Feature Selection: Understanding the correlations helps in selecting the most relevant features for predictive modeling. Features highly correlated with the target variable (electricity generation) are likely to contribute significantly to the model's accuracy.

3. Multicollinearity Detection: High correlations between predictor variables (independent variables) can indicate multicollinearity, a situation where variables are highly correlated, making it challenging to identify the individual effect of each variable on the target. Addressing multicollinearity is crucial for the accuracy and reliability of predictive models.

4. Insightful Visualization: Heatmaps provide a visually intuitive way to grasp complex relationships within the dataset, making it easier to communicate findings to stakeholders who might not be well-versed in statistical analysis.

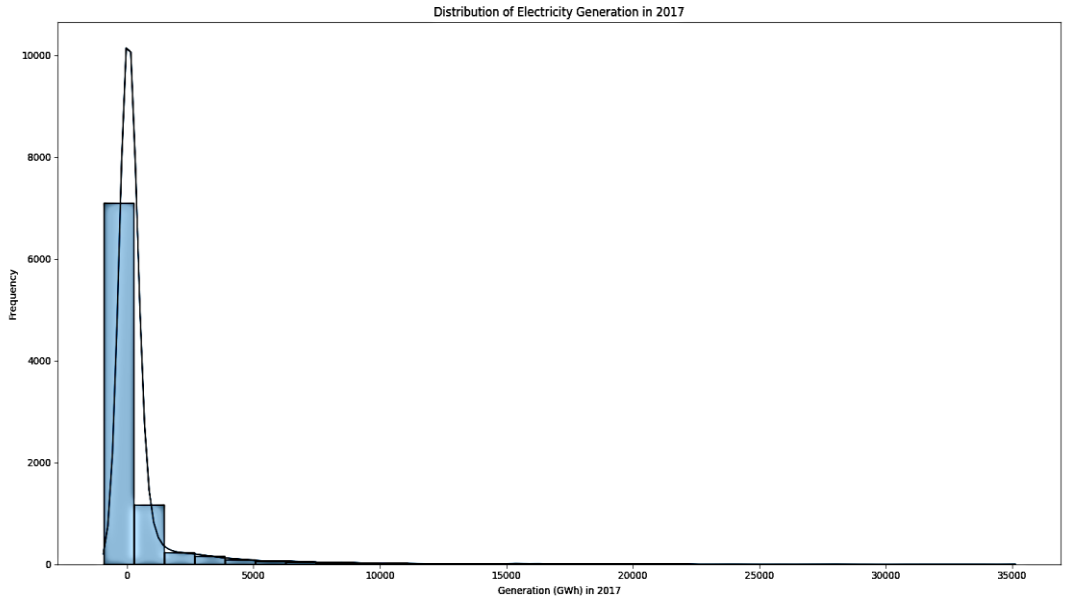

Figure 3. Distribution of Electrical Generation in 2017

The figure. 3 illustrates the electrical generation that was distributed in the year 2017 from the renewable energy sources.

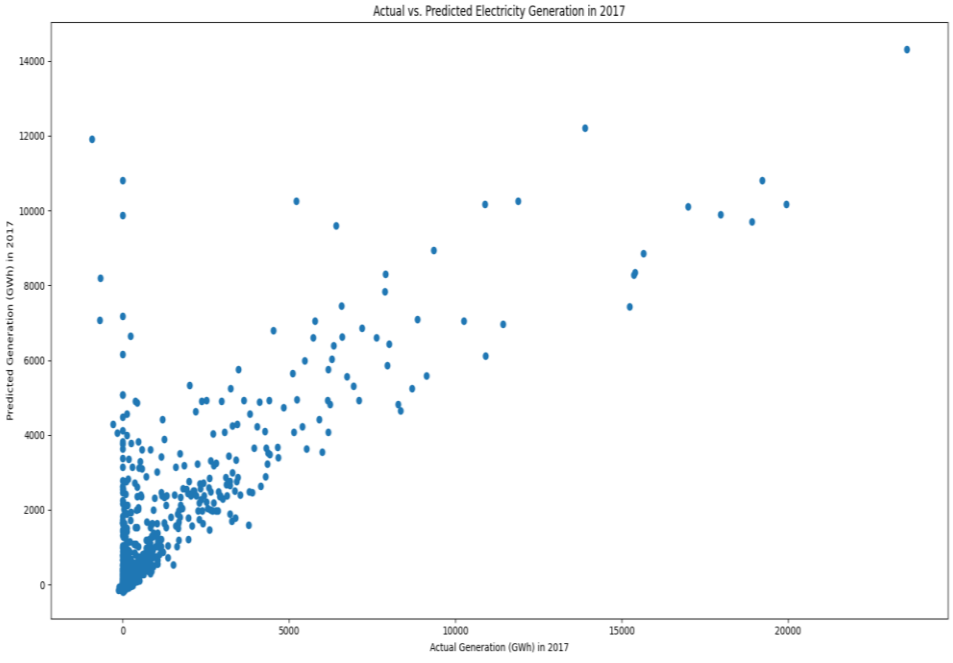

Figure 4. Actual vs Predicted Electricity Distribution

The output observed in the figure.4 shows the Actual vs Predicted Electricity Distribution

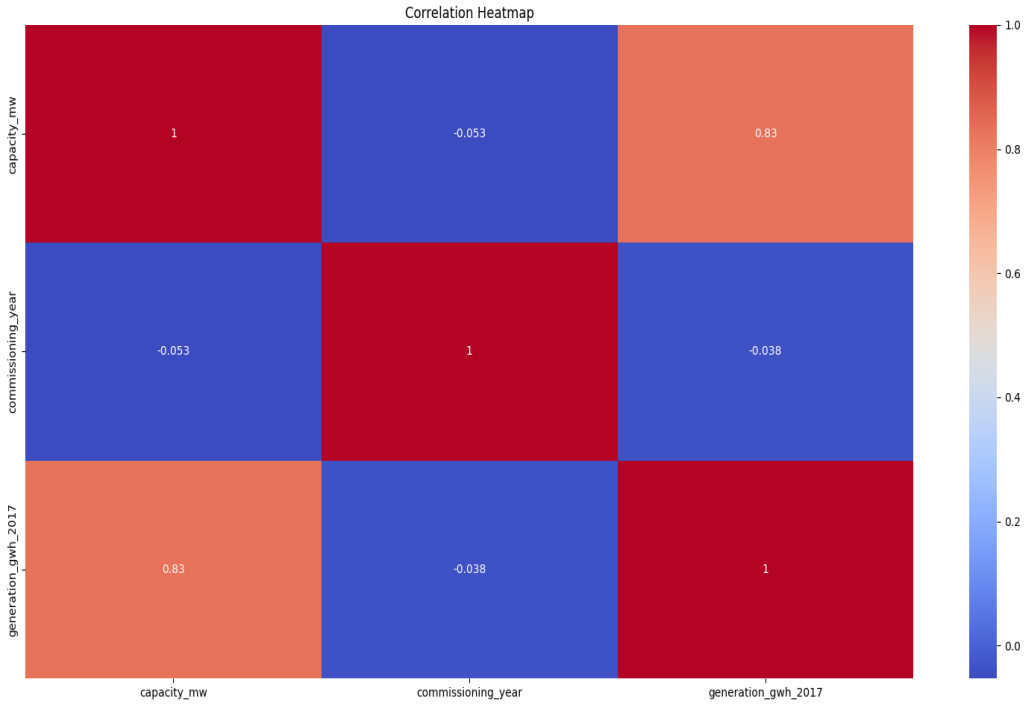

Figure 5. Correlation Heatmap

5. Results relation to the Energy Generation Prediction[1], [15], [16]

After obtaining the performance metric (Mean Squared Error) for the model's predictions, it's important to interpret this result in the context of the energy generation prediction. Here's how you can interpret the Mean Squared Error (MSE) in relation to the energy generation prediction:

Mean Squared Error (MSE)

Interpretation: The Mean Squared Error (MSE) measures the average squared difference between the predicted energy generation values and the actual energy generation values.

Relation to Energy Generation Prediction

A Lower MSE Indicates Better Prediction: A lower MSE indicates that, on average, the model's predictions are closer to the actual energy generation values. In other words, a smaller MSE suggests that the model is making more accurate predictions regarding energy generation.

Contextual Interpretation

High MSE (Large Errors): If the MSE is high, it means that the model's predictions have a significant amount of error concerning the actual energy generation. This might imply that the model's performance is not satisfactory, and the predictions can deviate substantially from the real values.

Low MSE (Small Errors): A low MSE indicates that the model's predictions closely match the actual energy generation. Smaller errors mean that the model is making precise predictions, capturing the underlying patterns and trends in the data related to energy generation.

Decision Making: Decision-makers and stakeholders can rely on a model with a low MSE for making informed decisions. For example, in the context of energy generation, accurate predictions are crucial for managing resources, optimizing energy distribution, and planning for future energy needs.

Improvement: If the MSE is higher than desired, it indicates room for improvement. Strategies such as feature engineering, exploring different algorithms, or collecting additional relevant data could be considered to enhance the model's accuracy and reduce prediction errors.

6. Conclusion

In this research the objective was to conduct a comprehensive analysis of the global power plant dataset, focusing on understanding the factors influencing electricity generation and building a predictive model for electricity generation in 2017. The analysis involved data cleaning, exploratory data analysis, and predictive modelling using linear regression.

(i) Data Cleaning and Exploration: The initial step involved rigorous data cleaning to ensure the dataset's integrity. Rows with missing values in critical columns

('capacity_mw', 'commissioning_year', 'generation_gwh_2017') were removed, establishing a reliable dataset. Exploratory data analysis techniques, including correlation heatmaps, were employed. These visualizations shed light on the relationships between key numerical variables. Notably, the analysis revealed the significance of 'capacity_mw' and 'commissioning_year' in influencing electricity generation.

(ii) Predictive Modelling: The predictive modelling phase employed a linear regression approach. The chosen features for prediction were 'capacity_mw' and 'commissioning_year,' aiming to forecast 'generation_gwh_2017.' The dataset was split into training and testing sets, facilitating an effective evaluation of the model. Mean squared error (MSE) was employed as a metric to gauge the model's performance. A lower MSE indicated higher predictive accuracy, signifying the importance of the selected features in determining electricity output.

(iii) Insights and Implications: The outcomes of the analysis carried vital implications. Policymakers could utilize the insights to encourage the development of more efficient, modern power plants. For investors, understanding the pivotal role of newer plants and increased capacity provided guidance for strategic investments, potentially resulting in higher electricity yields. Despite these insights, the study had limitations, such as the narrow focus on specific features. To address this, future work could expand the analysis to incorporate additional variables, including geographical factors and fuel types. Advanced machine learning techniques could also be integrated for a more nuanced understanding of electricity generation trends in the global power sector.

References

[1] N. Munier, “Transition to Renewable Energy: An Attempt to Model the Mix of Existing and Future Generation Technologies for 2035 and 2050,” *Biophysical Economics and Sustainability*, vol. 8, no. 4, p. 6, 2023, doi: 10.1007/s41247-023-00114-8.

[2] M. Imran, U. Zaman, Imran, J. Imtiaz, M. Fayaz, and J. Gwak, “Comprehensive survey of iot, machine learning, and blockchain for health care applications: A topical assessment for pandemic preparedness, challenges, and solutions,” *Electronics (Switzerland)*, vol. 10, no. 20. MDPI, Oct. 01, 2021. doi: 10.3390/electronics10202501.

[3] B. S. Neyigapula, “Synergistic Integration of Blockchain and Machine Learning: Advancements, Applications, and Challenges.” [Online]. Available: https://ssrn.com/abstract=4538519

[4] A. Qazi *et al.*, “Towards Sustainable Energy: A Systematic Review of Renewable Energy Sources, Technologies, and Public Opinions,” *IEEE Access*, vol. 7, pp. 63837–63851, 2019, doi: 10.1109/ACCESS.2019.2906402.

[5] M. J. B. Kabeyi and O. A. Olanrewaju, “Sustainable Energy Transition for Renewable and Low Carbon Grid Electricity Generation and Supply,” *Frontiers in Energy Research*, vol. 9. Frontiers Media S.A., Mar. 24, 2022. doi: 10.3389/fenrg.2021.743114.

[6] M. Zhou, “Exploring Application of Machine Learning to Power System Analysis.” [Online]. Available: www.interpss.org.

[7] A. Kumbhar, P. G. Dhawale, S. Kumbhar, U. Patil, and P. Magdum, “A comprehensive review: Machine learning and its application in integrated power system,” *Energy Reports*, vol. 7, pp. 5467–5474, 2021, doi: https://doi.org/10.1016/j.egyr.2021.08.133.

[8] R. Vaish, U. D. Dwivedi, S. Tewari, and S. M. Tripathi, “Machine learning applications in power system fault diagnosis: Research advancements and perspectives,” *Eng Appl Artif Intell*, vol. 106, p. 104504, 2021, doi: https://doi.org/10.1016/j.engappai.2021.104504.

[9] A. Entezari, A. Aslani, R. Zahedi, and Y. Noorollahi, “Artificial intelligence and machine learning in energy systems: A bibliographic perspective,” *Energy Strategy Reviews*, vol. 45, p. 101017, 2023, doi: https://doi.org/10.1016/j.esr.2022.101017.

[10] L. Wehenkel, “Machine-learning approaches to power-system security assessment,” *IEEE Expert-Intelligent Systems and their Applications*, vol. 12, no. 5, pp. 60–72, Sep. 1997, doi: 10.1109/64.621229.

[11] F. Chen, H. Wan, H. Cai, and G. Cheng, “Machine Learning in/for Blockchain: Future and Challenges,” Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.06189

[12] “ENERGYDATA.INFO.”

[13] S. Zhao, Y. Guo, Q. Sheng, and Y. Shyr, "Advanced Heat Map and Clustering Analysis Using Heatmap3," *Biomed Res Int*, vol. 2014, 2014, doi: 10.1155/2014/986048.

[14] B. C. M. (Benno) Haarman, R. F. Riemersma-Van der Lek, W. A. Nolen, R. Mendes, H. A. Drexhage, and H. Burger, "Feature-expression heat maps – A new visual method to explore complex associations between two variable sets," *J Biomed Inform*, vol. 53, pp. 156–161, 2015, doi: https://doi.org/10.1016/j.jbi.2014.10.003.

[15] S. Islam, "Impact investing in social sector organisations: a systematic review and research agenda," *Accounting and Finance*, vol. 62, pp. 709–737, Jan. 2022, doi: 10.1111/acfi.12804.

[16] R. K. F. Bresser and C. Powalla, "Practical implications of the resource-based view," *Zeitschrift für Betriebswirtschaft*, vol. 82, no. 4, pp. 335–359, 2012, doi: 10.1007/s11573-012-0553-4.