



Comparative Analysis of Machine Learning Models for Daytime Power Generation Prediction

Marwa M. Eid ¹*, Anis Ben Ghorbal ²

¹ Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura 11152, Egypt

² Department of Mathematics and Statistics, Faculty of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia

Emails: mmm@ieee.org; assghorbal@imamu.edu.sa,

Abstract

This paper proposes to evaluate how different machine learning techniques can be used to predict daytime power generation based on the "Daily Power Generation Data" dataset. As a result of six models, which contain Random Forest Regressor, Decision Tree Regressor, Nearest Neighbors, Linear Regression, MLP Regressor, and SVR, a clear understanding has been accomplished by assessing the performance using multiple metrics. First of all, the Random Forest Regressor turned out to be the best in terms of the Mean Squared Error (MSE) of 3.57×10^{-6} , which was the lowest among the three ML models. The introduction of the paper highlights the role of precise planning of the power market and the consecutive sections describing the topic mathematically. The table below, with a total list of performance issues, explains why the Random Forest Regressor is the superior full-proof model using the lowest MSE, highest explained variance, and great resistance to outlying samples. The paper thus gave various useful approval criteria that, to a great extent, we can choose the best model out of them because the Random Forest Regressor was in a position to get the highest performance metrics.

Keywords: Power Generation; Daily Power Generation; Machine Learning; Random Forest Regressor.

1 Introduction

The ever-increasing global energy demand, coupled with the mounting concerns over environmental sustainability, has catalyzed a profound transformation in the energy landscape. This dual challenge has intensified the need for innovative, efficient, and sustainable solutions in power generation [1]. With global populations expanding and industrial activities accelerating, the world faces a pressing need to ensure that energy production keeps pace while mitigating its environmental footprint. The optimization of power generation processes has thus become a cornerstone in addressing these multifaceted challenges, as it plays a pivotal role in ensuring energy security, economic stability, and ecological preservation [2].

For decades, fossil fuels such as coal, oil, and natural gas have dominated the global energy mix, supplying the majority of the energy required to power economies. However, this reliance on nonrenewable resources has come at a significant cost. Fossil fuels are finite and unevenly distributed, leading to geopolitical vulnerabilities and market instabilities. Furthermore, their extraction, processing, and combustion contribute to environmental degradation, including air pollution, habitat destruction, and the release of greenhouse gases that drive climate change. These issues underscore the urgent need to transition towards cleaner, renewable, and more sustainable energy alternatives.

In response to these challenges, the optimization of power generation has emerged as a critical area of research. This optimization involves not only maximizing the efficiency of energy production but also minimizing environmental impacts through the adoption of innovative technologies and sustainable practices [3]. By exploring strategies to harness renewable energy sources, improve energy storage systems, and enhance grid management, researchers aim to bridge the gap between energy demand and environmental stewardship. The present study is grounded in this pursuit, seeking to advance knowledge and provide actionable insights for a more sustainable energy future.

This research is guided by three core objectives that address the pressing issues in contemporary power generation. Firstly, it seeks to identify and analyze the challenges inherent in existing power generation methods, particularly those that hinder sustainability and efficiency. This includes examining the technical, economic, and policy-related barriers that impede progress. Secondly, the study investigates emerging technologies and innovations—ranging from advanced machine learning algorithms for power forecasting to novel renewable energy systems—that hold the potential to transform the energy sector. Finally, it explores strategies to reduce the environmental footprint of power generation, emphasizing approaches that prioritize ecological balance, resource conservation, and long-term sustainability [4, 5].

The multidisciplinary nature of this study is a key strength, as it draws on expertise from engineering, environmental science, economics, and policy-making to develop holistic solutions. By fostering collaboration among stakeholders, including researchers, industry leaders, and policymakers, the study aims to create a shared vision for sustainable energy systems. This collaborative approach not only enhances the feasibility of proposed solutions but also ensures their alignment with societal needs and values.

As the world grapples with the challenges of climate change and resource scarcity, this research highlights the importance of transitioning to sustainable energy systems. It emphasizes the role of informed decision-making, technological innovation, and community engagement in achieving this transition. The insights provided in this paper aim to inspire action, drive innovation, and contribute to the global effort to build a resilient and sustainable energy infrastructure.

In the subsequent sections, this paper delves into the intricacies of current power generation methods, evaluates the potential of state-of-the-art technologies, and proposes actionable strategies for sustainable energy production. By addressing these critical issues, the study seeks to empower stakeholders with the knowledge and tools needed to navigate the complexities of the energy transition. Ultimately, this work aspires to contribute to a future where energy is abundant, accessible, and environmentally responsible, ensuring the well-being of both present and future generations [6, 7].

2 Literature Review

Power generation is an essential cornerstone of modern society, underpinning industries, infrastructure, and the daily lives of billions of people worldwide. As energy demand continues to rise due to population growth, urbanization, and industrial expansion, the need for efficient, sustainable, and reliable power generation has become more pressing than ever. The dual challenges of meeting these growing demands while minimizing environmental impacts have catalyzed a wave of innovation and research aimed at optimizing energy systems. This literature review delves into traditional and renewable energy sources, advances in smart grid technologies, and cutting-edge methodologies for enhancing power generation efficiency and resilience.

Renewable energy sources, such as solar, wind, and hydropower, have emerged as critical components of the global energy transition. Among these, solar energy is particularly prominent due to its abundance and potential for widespread adoption. However, the efficiency of solar energy systems heavily depends on accurate solar radiation estimation and system optimization. Research studies such as [8] have introduced innovative hybrid modeling techniques, including the PSO-ELM (Particle Swarm Optimization-Extreme Learning Machine) model, which has demonstrated significant improvements in predicting daily solar radiation, especially in regions lacking extensive onsite data.

Similarly, power prediction methodologies for photovoltaic (PV) plants have been enhanced through advanced machine learning algorithms. For instance, [9] highlights the application of sophisticated neural networks to

improve the accuracy and reliability of PV power forecasting, addressing challenges such as weather variability and data scarcity. Concurrently, [10] proposed the SDA-GA-ELM (Stacked Denoising Autoencoder-Genetic Algorithm-Extreme Learning Machine) hybrid model, specifically designed for precise hourly predictions of PV power output. This innovative approach underscores the critical role of accurate forecasting in optimizing solar energy systems and integrating them into the broader energy grid.

The transition to smart grid technologies represents a paradigm shift in energy management, enabling greater efficiency, resilience, and sustainability. However, this transition is not without challenges, including demand-side management, cybersecurity, and grid infrastructure optimization. The study conducted in [11] provides valuable insights into the application of machine learning algorithms for smart grid technologies, emphasizing their potential to enhance data processing, optimize resource allocation, and improve network reliability.

Additionally, the LSTM-PC (Long Short-Term Memory Predictive Control) model introduced in [12] showcases the potential of advanced deep learning techniques in addressing the inherent unpredictability of renewable energy sources. This model is particularly effective in forecasting energy production from PV plants, offering solutions to challenges such as fluctuating energy demand and weather-dependent variability.

Reducing greenhouse gas emissions remains a central goal in the transition to sustainable energy systems. Research such as [13] employs state-of-the-art deep learning models to predict emissions from Turkey's electricity sector, highlighting the urgent need to curtail fossil fuel consumption. By combining traditional methodologies with modern computational approaches, this study underscores the critical importance of transitioning to cleaner energy sources to mitigate environmental degradation.

Power system operations are inherently complex, requiring robust strategies to manage uncertainty and variability. Innovative solutions, such as the scenario generation methods proposed in [14], provide statistically valid projections for power system management, facilitating more informed decision-making. Similarly, the data-driven framework developed in [15] leverages advanced techniques like STL (Seasonal and Trend decomposition using Loess) and LSTM to improve the accuracy of monthly renewable energy forecasts, ensuring grid stability and reliability.

Ensuring grid observability is critical for the efficient operation of modern electricity systems. Studies such as [16] explore the application of state estimation approaches and machine learning techniques to optimize grid operations under varying conditions. These methodologies enable real-time status predictions, enhance system resilience, and pave the way for adaptive energy management strategies that respond dynamically to evolving energy demands and environmental challenges.

In summary, the literature underscores the necessity of adopting clean, sustainable, and innovative technologies within the power generation sector. The reviewed studies highlight the transformative potential of renewable energy sources, advanced forecasting methodologies, and smart grid technologies in addressing contemporary energy challenges. Furthermore, the importance of interdisciplinary collaboration among researchers, policymakers, and industry stakeholders is emphasized as a means of fostering the development of resilient and environmentally friendly energy systems. As the global energy landscape continues to evolve, these advancements lay the groundwork for achieving a sustainable and equitable energy future.

3 Dataset

3.1 Dataset Description

The "Daily Power Generation Data" dataset [17] is notable for its comprehensive information on power generation activities, offering insights sorted by geographical stations. Covering a significant period from September 1, 2017, to January 19, 2023, this dataset provides a detailed account of electricity production history, enabling a deep understanding of power generation dynamics over time.

However, it is essential to highlight significant data gaps on specific dates, including October 2, 2017, November 19, 2017, November 26, 2017, April 3, 2018, and April 4, 2018, as well as the prolonged gap from March

19, 2020, to May 31, 2020. These gaps obscure critical insights into power generation during these periods, presenting challenges for comprehensive analysis and decision-making processes.

Furthermore, the dataset is available in two distinct file formats, necessitating precise processing and analysis to ensure consistency and accuracy across both formats. Overcoming these challenges is paramount for reliable data preparation, particularly for predictive analysis and informed decision-making in energy management. Exploring strategic approaches to address data gaps and associate power generation processes with level estimations will be essential to enhance the dataset's trustworthiness and usefulness for future analysis. By confronting these challenges head-on, researchers can unlock valuable insights and facilitate informed decision-making in the realm of power generation dynamics.

3.2 Dataset Preprocessing Steps

Before commencing the analysis of daily power generation data, a meticulous preparatory phase was undertaken to ensure the dataset's integrity and consistency. The subsequent processes delineate the preprocessing [18] methods employed:

Handling Missing Data: The identification and resolution of dates with missing data involved a thorough review to understand the underlying reasons. Imputation techniques were then applied to fill these gaps, offering researchers flexibility in choosing methods tailored to the data's nature and analytical objectives. Techniques such as mean imputation, interpolation, or predictive modeling were considered based on the context of the missing data and the research objectives.

File Format Harmonization: Given the disparity in file formats across the dataset, a harmonization method was implemented to ensure accurate representation of the dataset. This involved converting and aligning the storage format to establish a well-ordered structure conducive to future analyses. Additionally, efforts were made to reconcile any discrepancies in data encoding or formatting conventions between the two files to ensure seamless integration and consistency.

Standardization of Units: A systematic standardization procedure was executed on the dataset's units of measurement to enhance clarity and facilitate comparisons across variables. Specifically, a methodical transformation from megaunits (MU) to megawatts (MW) was conducted, promoting a cohesive and uniform analysis. This conversion facilitated a more intuitive understanding of power generation metrics and eliminated potential confusion arising from disparate unit representations.

Ensuring dataset integrity was crucial during preparation, with subsequent verifications conducted to address any inconsistencies. Through cross-variable consistency tests, researchers scrutinized relationships between variables, rectifying any discrepancies detected to ensure coherence and accuracy.

3.3 Descriptive Statistics

Descriptive statistics were instrumental in comprehensively understanding the traits and patterns within the "Daily Power Generation Data," offering a quantitative overview of key characteristics and scrutinizing the dataset's major tendencies and variation. Spanning from September 1, 2017, to January 19, 2023, the dataset captures a rich tapestry of temporal dynamics in daily power generation. Across diverse power stations, each contributing uniquely to the overall landscape, computed descriptive statistics shed light on individual performance metrics.

The histogram in Figure 1 vividly displays variations between projected and actual power generation, using bars to represent real data.

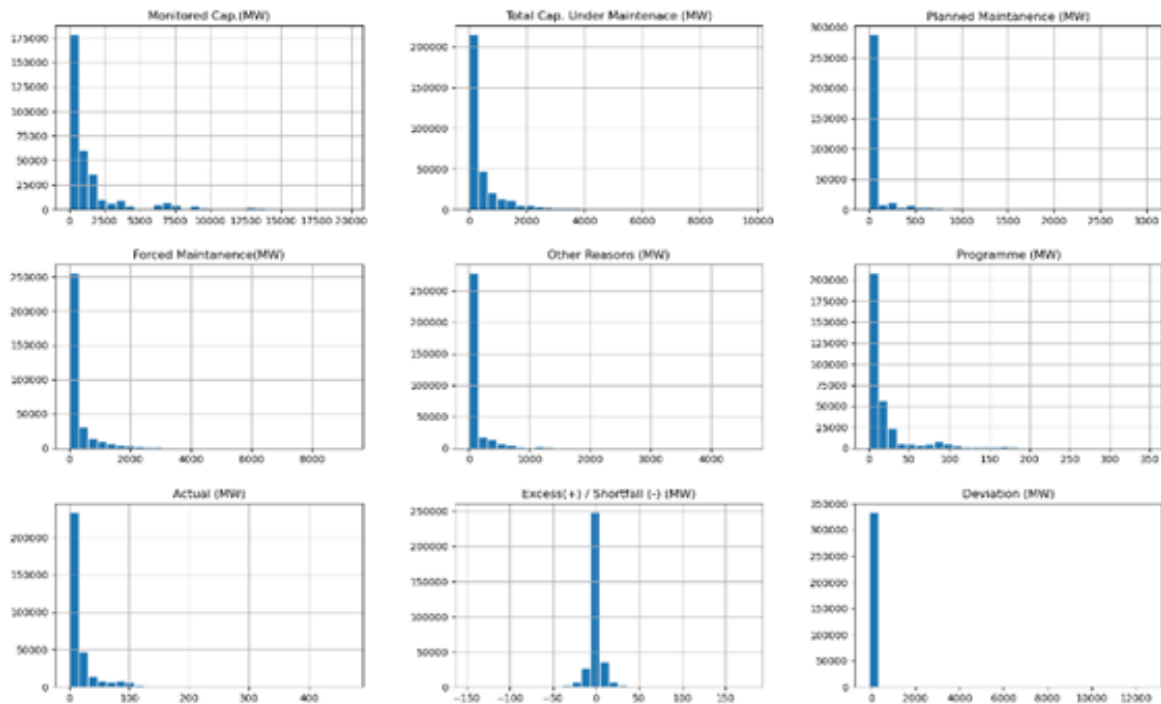


Figure 1: Histogram illustrating variations between projected and actual power generation.

4 Results

This section is devoted to the outcomes of applying several machine learning models to the "Daily Power Output Data" dataset. The models were implemented to forecast power generation from various power stations, using features such as Expected Power Generation, Actual Power Generation, Deviation, and others. The objective of this study is to evaluate the models and determine the one that provides the best prediction performance.

The following machine learning models were evaluated:

1. Random Forest Regressor
2. Decision Tree Regressor
3. Nearest Neighbors
4. Linear Regression
5. MLP Regressor
6. SVM Regressor

For each model, a set of performance metrics was developed to comprehensively assess the model's accuracy in forecasting power production. The metrics considered include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Explained Variance Score (EVS), Max Error, Median Absolute Error (MedAE), Mean Absolute Percentage Error (MAPE), R-squared (R²), Modified Tchebycheff Distance (MTD), Relative Root Mean Squared Error (RRMSE), Willmott Index, Mean Bias Error (MBE), and Standard Deviation (SD).

The following sections provide detailed results for each model and their performance in predicting power generation.

4.1 Machine Learning Models

In this section, we detail the machine learning models applied to the "Daily Power Generation Data" dataset. These models represent diverse approaches to predictive analytics, with varying design features and mathematical foundations.

4.1.1 Random Forest Regressor

Explanation: The Random Forest Regressor is an ensemble learning method that combines multiple decision trees, each trained on a different data subset. The final prediction is the average of individual tree predictions, reducing overfitting and improving accuracy.

Mathematical Foundation:

$$\hat{y} = \frac{1}{T} \sum_{i=1}^T h_i(x),$$

where $h_i(x)$ represents the prediction of the i -th tree, and T is the total number of trees.

4.1.2 Decision Tree Regressor

Explanation: The Decision Tree Regressor splits data into subsets at each node based on the most important feature. Final predictions are the mean values of target variables at the leaf nodes.

Mathematical Foundation:

$$\hat{y} = \frac{1}{|R_j|} \sum_{x_i \in R_j} y_i,$$

where R_j is the region represented by the j -th leaf node, and y_i are the target values within R_j .

4.1.3 Nearest Neighbors

Explanation: The k-Nearest Neighbors algorithm predicts a target value based on the average of the k -nearest neighbors in the feature space.

Mathematical Foundation:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^k y_i,$$

where y_i represents the target values of the k -nearest neighbors.

4.1.4 Linear Regression

Explanation: Linear regression models the relationship between the target variable and independent variables as a linear function.

Mathematical Foundation:

$$\hat{y} = X\theta + b,$$

where X are the features, θ are the weights, and b is the bias term.

4.1.5 MLP Regressor (Multi-Layer Perceptron)

Explanation: The MLP Regressor is a neural network model with input, hidden, and output layers. It captures complex, nonlinear relationships between variables.

Mathematical Foundation:

$$\hat{y} = f(W^T X + b),$$

where W represents the weights, X the input features, and f the activation function.

4.1.6 SVM Regressor (Support Vector Regressor)

Explanation: SVR fits a nonlinear hyperplane in a transformed feature space to minimize the total deviation from the observed data.

Mathematical Foundation:

$$\hat{y} = \sum_{i=1}^n \alpha_i K(x, x_i) + b,$$

where K is the kernel function, α_i are the support vector weights, and b is the bias.

4.2 Performance Metrics

Key Metrics:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Explained Variance Score (EVS)
- Max Error
- Median Absolute Error (MedAE)
- Mean Absolute Percentage Error (MAPE)
- R-squared (R²)
- Modified Tchebycheff Distance (MTD)
- Relative Root Mean Squared Error (RRMSE)
- Willmott Index
- Mean Bias Error (MBE)
- Standard Deviation (SD)

Each metric evaluates a specific aspect of model performance, providing a well-rounded analysis of the model's predictive capabilities.

4.3 Results

Table 1 presents the performance metrics for the evaluated machine learning models.

Table 1: Regression Results of Machine Learning Models

Model	MSE	RMSE	MAE	EVS	Max Error	MedAE	MAPE	R2	MTD	RRMSE	Willmott Index	MBE
Random Forest Regressor	3.57E-06	0.018	0.007	0.804	0.043	0.007	0.35	0.99	0.018	0.003	0.99	0.0
Decision Tree Regressor	6.16E-06	0.024	0.010	0.214	0.219	0.005	0.58	0.892	0.024	0.010	0.81	0.0
Nearest Neighbors	8.51E-06	0.029	0.004	0.750	0.601	0.002	0.743	0.604	0.029	0.011	0.69	0.0
Linear Regression	1.37E-05	0.036	0.018	0.812	0.655	0.007	0.804	0.301	0.036	0.013	0.71	0.0
MLP Regressor	1.46E-05	0.038	0.020	0.107	0.554	0.008	0.619	0.195	0.038	0.016	0.59	0.0
SVR	8.1E-05	0.090	0.025	-0.130	0.904	0.037	1.502	0.233	0.090	0.050	0.33	0.0

4.4 Discussion of Key Findings

- **Random Forest Regressor:** Achieved the best performance, with the lowest MSE, RMSE, and MAE, along with high EVS and R2 values, indicating its suitability for power generation prediction.
- **Decision Tree Regressor:** Exhibited good performance, with moderate accuracy metrics, slightly less effective than Random Forest.
- **Nearest Neighbors:** Showed moderate prediction accuracy but struggled to capture complex patterns.
- **Linear Regression:** Performed reasonably well, especially in capturing linear relationships.
- **MLP Regressor:** Demonstrated the ability to handle nonlinear patterns but was less effective than Random Forest.
- **SVR:** Performed poorly, with negative EVS and R2 values, indicating insufficient representation of the underlying data patterns.

The Random Forest Regressor achieved the best performance among all models and is recommended for predicting power generation based on the given dataset.

5 Conclusion

In summary, this research aims to achieve accurate forecasts of power generation using a data-driven approach accomplished through an in-depth analysis of the performance of multiple machine learning models in the context of the "Daily Power Generation Data" dataset. The exploration covered six models: Random Forest Regressor, Decision Tree Regressor, Nearest Neighbors, Linear Regression, MLP Regressor, and SVR, evaluated by performance metrics on all fronts.

Notably, the Random Forest Regressor emerged as the most effective model, achieving the lowest Mean Squared Error (MSE) of 3.57×10^{-6} , indicating superior predictiveness. The thorough evaluation of metrics such as RMSE, MAE, EVS, and Max Error further demonstrated the robustness of Random Forest Regression, showcasing its resistance to outliers and external variations in power data analysis.

The results obtained from this study hold significant value for stakeholders involved in power company management, providing actionable insights into the most effective method for implementing predictive analytics. Among the evaluated models, the Random Forest Regressor demonstrated dominant competence, with other models performing sufficiently but less effectively than the RF Regressor. This statistically validates that the RF Regressor is the most suitable model for accurate and reliable power generation predictions.

However, the development of an ultimate framework to address all target aspects and limitations remains a challenging process. Further advancements in model development and data analysis are encouraged to refine predictive accuracy and reliability for power generation applications.

References

- [1] E. M. Almetwally and M. A. Meraou. Application of environmental data with new extension of nadarajah-haghighi distribution. *Computational Journal of Mathematical and Statistical Sciences*, 1(1):26–41, 2022.
- [2] H. Z. Muhammed and E. M. Almetwally. Bayesian and non-bayesian estimation for the shape parameters of new versions of bivariate inverse weibull distribution based on progressive type ii censoring. *Computational Journal of Mathematical and Statistical Sciences*, 3(1):85–111, 2024.
- [3] Khader M. Hamdia, Xiaoying Zhuang, and Timon Rabczuk. An efficient optimization approach for designing machine learning models based on genetic algorithm. *Neural Computing and Applications*, 33(6):1923–1933, March 2021.
- [4] Yilin Ma, Ruizhu Han, and Weizhong Wang. Portfolio optimization with return prediction using deep learning and machine learning. *Expert Systems with Applications*, 165:113973, March 2021.
- [5] R. Alotaibi, G. R. AL-Dayian, E. M. Almetwally, and H. Rezk. Bayesian and non-bayesian two-sample prediction for the fréchet distribution under progressive type ii censoring. *AIP Advances*, 14(1):015137, 2024.
- [6] Jichao Li, Xiaosong Du, and Joaquim R. R. A. Martins. Machine learning in aerodynamic shape optimization. *Progress in Aerospace Sciences*, 134:100849, October 2022.
- [7] Defu Zhu, Biaobiao Yu, Deyu Wang, and Yujiang Zhang. Fusion of finite element and machine learning methods to predict rock shear strength parameters. *Journal of Geophysics and Engineering*, 21(4):1183–1193, August 2024.
- [8] Vikram Pasupuleti, Bharadwaj Thuraka, Chandra Shikhi Kodete, and Saiteja Malisetty. Enhancing Supply Chain Agility and Sustainability through Machine Learning: Optimization Techniques for Logistics and Inventory Management. *Logistics*, 8(3):73, September 2024. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
- [9] A. Djaafari, A. Ibrahim, N. Bailek, K. Bouchouicha, and M. A. Hassan. Hourly predictions of direct normal irradiation using an innovative hybrid lstm model for concentrating solar power projects in hyper-arid regions. *Energy Reports*, 8:15548–15562, 2022.
- [10] Masoud Aliramezani, Charles Robert Koch, and Mahdi Shahbakhti. Modeling, diagnostics, optimization, and control of internal combustion engines via modern machine learning techniques: A review and future directions. *Progress in Energy and Combustion Science*, 88:100967, January 2022.
- [11] Heng Chi, Yuyu Zhang, Tsz Ling Elaine Tang, Lucia Mirabella, Livio Dalloro, Le Song, and Glaucio H. Paulino. Universal machine learning for topology optimization. *Computer Methods in Applied Mechanics and Engineering*, 375:112739, March 2021.
- [12] Kerr Ding, Michael Chin, Yunlong Zhao, Wei Huang, Binh Khanh Mai, Huanan Wang, Peng Liu, Yang Yang, and Yunan Luo. Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering. *Nature Communications*, 15(1):6392, July 2024. Publisher: Nature Publishing Group.
- [13] Ponnarasan Krishnan. Ai-Driven Optimization In Healthcare: Machine Learning Models For Predictive Diagnostics And Personalized Treatment Strategies. *Well Testing Journal*, 33(S2):10–33, September 2024. Number: S2.
- [14] S. Bhavsar, R. Pitchumani, and M. A. Ortega-Vazquez. Machine learning enabled reduced-order scenario generation for stochastic analysis of solar power forecasts. *Applied Energy*, 293:116964, 2021.
- [15] S. Ding, H. Zhang, Z. Tao, and R. Li. Integrating data decomposition and machine learning methods: an empirical proposition and analysis for renewable energy generation forecasting. *Expert Systems with Applications*, 204:117635, 2022.
- [16] D. Mukherjee, S. Chakraborty, and S. Ghosh. Power system state forecasting using machine learning techniques. *Electrical Engineering*, 104(1):283–305, 2022.

- [17] Daily power generation data. <https://www.kaggle.com/datasets/arvindnagaonkar/power-generation-data>, 2024. Accessed: Mar. 13, 2024.
- [18] S. Alam and N. Yao. The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. *Computational and Mathematical Organization Theory*, 25(3):319–335, 2019.