Advanced Big Data Analytics Framework for Predicting Renewable Energy Production

Tapan Patel
Pandit Deendayal Energy University
School of Technology
Computer Science and Engineering
Gandhinagar, Gujarat India
tapan.patel2003@gmail.com

Shlok Sanghvi
Pandit Deendayal Energy University
School of Technology
Computer Science and Engineering
Gandhinagar, Gujarat, India
shlok.sanghvi06783@gmail.com

Harit Dobariya
Pandit Deendayal Energy University
School of Technology
Computer Science and Engineering
Gandhinagar, Gujarat, India
haritcodes@gmail.com

***Abstract*—This paper addresses the critical need for accurate prediction of renewable energy production, focusing on solar and wind power systems. Accurate forecasts are essential for optimizing energy generation, resource allocation, and long-term planning. Existing prediction models often face challenges due to environmental variability and complexity. This study proposes an advanced big data analytics framework tailored for precise renewable energy production prediction, aiming to address these challenges. Literature review showcases various studies on renewable energy prediction, highlighting the significance of big data analytics and machine learning in addressing forecasting challenges. Data analysis provides insights into the correlation between environmental factors and energy output. The study implements linear regression and decision tree models to predict solar and wind energy production, discussing challenges and strategies for improvement. Results underscore the importance of refining modeling strategies for enhanced prediction accuracy. The paper concludes by emphasizing the necessity of advanced analytics in optimizing renewable energy production.**

I. Introduction

The global energy landscape is rapidly evolving as societies worldwide seek to transition towards more sustainable and environmentally friendly energy sources. Among these, solar and wind power have emerged as key players, offering abundant, renewable, and clean energy potential. However, the effective integration of these renewable sources into the energy grid necessitates accurate prediction of their energy output. Accurate forecasts of renewable energy production are vital for efficient resource management, grid stability, and informed investment decisions. Yet, existing prediction models face significant challenges due to the inherent variability and complexity of environmental factors. Fluctuations in solar irradiance, wind patterns, and weather dynamics present obstacles to precise forecasting.

This study proposes an advanced big data analytics framework tailored for precise renewable energy production prediction. By leveraging the power of big data, machine learning, and statistical modeling, this framework aims to enhance the accuracy and reliability of energy forecasts. Through the integration of high-resolution weather data and advanced analytical techniques, the proposed framework seeks to overcome the limitations of traditional forecasting methods and enable more effective management of renewable energy resources. In the subsequent sections, we will delve into the intricacies of renewable energy prediction, review relevant literature, outline the methodology employed in this study, present empirical findings, and discuss the implications of our research for the advancement of sustainable energy technologies and practices.

II. Problem Statement

The increasing reliance on renewable energy sources, such as solar and wind power, underscores the critical need for accurate prediction of energy output to ensure effective integration into the grid and sustainable energy management. Accurate forecasts play a pivotal role in facilitating efficient resource allocation, enhancing grid stability, and informing strategic investment decisions, thereby contributing significantly to the transition towards a low-carbon economy. However, existing prediction models encounter formidable challenges stemming from environmental variability, non-linear relationships, and data scarcity. These limitations impede the optimal utilization of renewable energy resources and present significant barriers to achieving energy transition goals.

To address these challenges, this study proposes an advanced big data analytics framework tailored specifically for precise prediction of renewable energy production. By harnessing the power of machine learning algorithms, sophisticated data fusion techniques, and high-resolution weather data, the framework aims to enhance the accuracy and reliability of energy forecasts for solar and wind power systems. Ultimately, the goal is to empower energy providers, policymakers, and investors with actionable insights to make informed decisions, optimize energy generation strategies, and expedite the transition to a sustainable energy future.

III. Literature Review

The study by Ceci et al. (2015) focuses on predicting renewable energy output across diverse locations, which is crucial for marketing and energy management purposes. Its objective is to design a Virtual Power Operating Center (Vi-POC) prototype for real-time energy prediction using advanced Big Data techniques. Utilizing a distributed system, the research employs HBase over Hadoop to store data from energy plants and weather services. Machine learning algorithms are applied to forecast PV energy production one day ahead, with initial results indicating Vi-POC's potential to enhance energy market strategies. Discussion highlights the system's fault tolerance and the separation of computation and storage. The study concludes by underscoring the need for further research to refine prediction techniques for wider commercial application.

Mujeeb et al. (2019) discuss the importance of accurate wind power forecasting, so that it can be easily integrated into existing systems. In the study, an Efficient Deep Convolution Neural Network (EDCNN) is employed for wind power forecasting. The model's performance was evaluated using data from a wind farm in Maine, USA. The proposed EDCNN outperforms typical CNN and SELU CNN models in wind power forecasting, as evidenced by lower MAPE, NRMSE, and MAE values. Statistical tests validate the significant improvement of EDCNN over the compared models. The proposed DSM algorithm effectively shifts loads to off-peak hours, reducing consumption costs. This paper provides a comprehensive approach to address challenges in power systems through accurate wind power forecasting and efficient demand-side management.

In the study focused on integrating variable renewable energy sources into utility operations, Haupt and Kosović (2017), aim to enhance forecast accuracy using a big data approach. They utilize a large dataset comprising real-time observations and various model outputs tailored for both day-ahead planning and real-time operations. Their methodology leverages computational intelligence to amalgamate data across temporal and spatial scales effectively. The results demonstrate improved forecast reliability, crucial for managing renewable energy variability. This outcome is discussed in terms of its implications for grid stability and utility operations, highlighting the critical role of advanced analytics in optimizing renewable integration. The conclusion emphasizes the necessity of big data techniques in overcoming challenges posed by renewable energy sources.

Buturache and Stancu's (2021), study addresses the prediction of wind energy output, crucial for minimizing operational costs and ensuring the safe operation of power grids. They employ a comprehensive dataset including historical wind patterns and turbine data, applying various machine learning models such as neural networks, support vector regression, and random forests. Their methodology involves a structured evaluation using the CRISP-DM framework, which aids in identifying the most effective predictive models. The results reveal that certain models, particularly neural networks, provide optimal performance in terms of accuracy and resource efficiency. The study discusses these outcomes with respect to their practical applicability in wind energy management, providing valuable insights for industry professionals. The conclusion highlights the potential of machine learning to enhance wind energy prediction and contribute to more sustainable energy practices.

Je et al. (2021) focus on improving the accuracy of demand forecasting for electric power production using a big data approach centered around a photovoltaic power plant. Their objective is to better align electricity demand and supply using big data analytics, employing a dataset that includes consumption patterns and production forecasts. The methodology utilizes game theory and big data virtualization techniques to predict power supply needs and integrate renewable energy sources into the grid efficiently. The results demonstrate enhanced forecasting accuracy and more efficient energy distribution, crucial for incorporating renewable energy into existing systems. The discussion explores the implications of these results for future energy management practices, emphasizing the role of big data in achieving a balanced and flexible power production system. The conclusion underscores the importance of innovative data-driven approaches in advancing the predictability and efficiency of renewable energy systems.

IV. Methodology

A. Data Collection and Data Gathering

The study utilized a substantial dataset from Chen and Xu [**?**], comprising historical records of solar irradiance, wind speed, temperature, humidity, and energy production metrics. The dataset underwent preprocessing steps to ensure data quality and uniformity, including filtering, cleaning, and normalization.

B. Data Analysis and Insights

Data analysis was conducted to identify correlations between environmental factors and energy output for solar and wind power systems. Diurnal and seasonal patterns in energy production were observed, along with the impact of meteorological parameters on power generation.

1) Solar Energy Production:

- **Observed Diurnal Generation Trends:** The line graphs demonstrate a clear diurnal pattern in power generation, with peak production aligning with solar noon. This evidences the critical influence of solar irradiance on energy yield, emphasizing the importance of sun-path tracking for solar installations.
- **Impact of Environmental Factors:** A correlation between air temperature and solar power output is observed, suggesting that warmer temperatures could be associated with higher solar irradiance and thus greater power generation. However, the relationship isn't necessarily linear, as extremely high temperatures can decrease the efficiency of photovoltaic cells.
- **Seasonal Power Output Variations:** The monthly power generation bar graphs exhibit a seasonal trend with heightened energy production in the summer months due to longer daylight hours and a decline in the winter months, indicating the necessity of energy storage solutions or complementary energy sources to counteract seasonal variability.
- **Correlational Dynamics with Environmental Parameters:** Heatmaps from the correlation matrix show a strong positive correlation between total, direct, and global horizontal solar irradiance with power output. Conversely, a negative correlation with relative humidity is observed,

suggesting that drier conditions are more conducive for solar energy production.

2) Wind Energy Production:

- **Diurnal and Seasonal Wind Power Profiles:** Line graph analysis indicates that wind power production is less subject to diurnal shifts but exhibits noticeable fluctuations, which could be attributed to meteorological changes. Seasonal bar graphs display variances in production, with peaks often occurring in seasons characterized by meteorological transitions, such as spring and autumn.
- **Wind Speed and Altitude Correlation to Energy Output:** Data shows that wind speed increases with altitude and significantly impacts power generation. Bar graphs categorizing wind speeds at different heights reveal that higher altitudes offer a more consistent and powerful wind resource, directly translating to increased power output.
- **Meteorological Parameter Impact:** The correlation matrix heatmap indicates that while air temperature and atmospheric pressure have a nominal direct effect on wind energy production, wind speed exhibits a strong positive correlation. This underscores the preeminence of wind speed as a predictor of power generation in wind farms.

C. Problem Solution

1) Challenges Addressed:

- **Data Preprocessing and Scaling:** Effective management of data scaling is crucial for accurate prediction of solar and wind energy outputs. Given the impact of various environmental factors, we employed StandardScaler to normalize the training and testing datasets, which ensures consistency across variable scales and enhances model accuracy.

2) Implemented Solutions and Results:

- **Linear Regression Models:** Implementation: Linear Regression models were used to predict energy outputs from solar and wind datasets, which had been scaled to handle variable disparities effectively.
  - **Solar Energy:** The Linear Regression model for solar data showed an MAE of 6.50 and an RMSE of 12.20. These metrics indicate the model's deviation from the actual data points, reflecting moderate prediction accuracy.
  - **Wind Energy:** For wind data, the model reported an MAE of 9.67 and an RMSE of 13.39, demonstrating a slightly higher error rate compared to the solar model, which suggests challenges in capturing the variability in wind energy data.
- **Decision Tree Regressor:** Implementation: Decision Tree Regressors were tested to address the non-linear relationships within the environmental data that Linear Regression models might miss.
  - **Challenges and Results:** Overfitting Issue: Despite their capability to model complex patterns, the Decision Tree models exhibited overfitting, performing well on training data but poorly on testing data.
  - **Strategies to Mitigate Overfitting:** To combat the overfitting observed in Decision Tree models, strategies such as model simplification (pruning, maximum depth), cross-validation, and ensemble methods like Random Forest or Gradient Boosting were considered. These methods are aimed at balancing bias and variance to improve model performance.

V. Results and Data Analysis and Insights

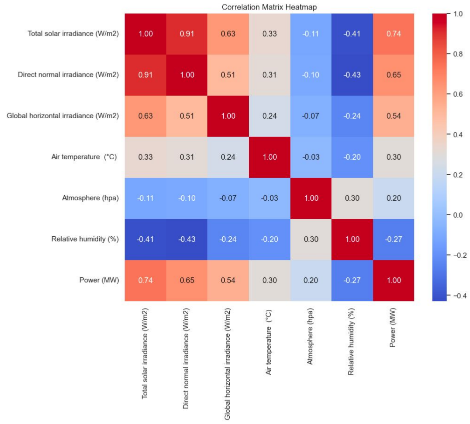

Fig. 1. Correlation matrix heatmap of solar power

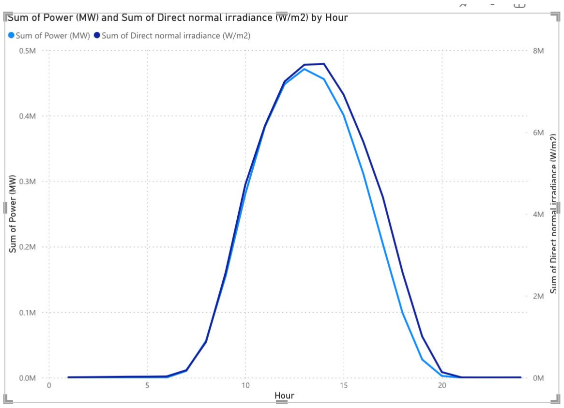

Fig. 2. Power vs. Normal Irradiance

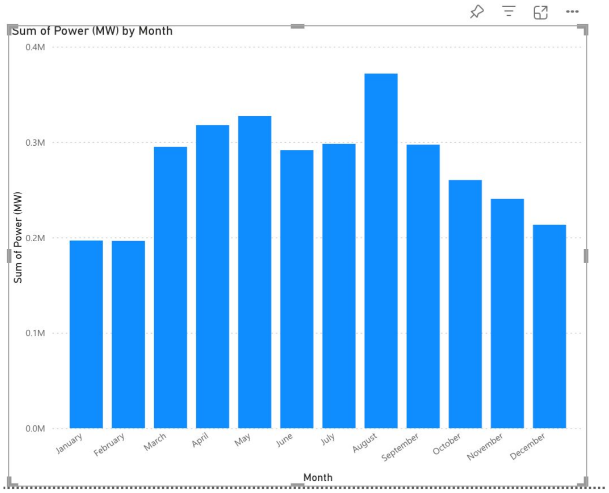

Fig. 3. Month wise Power Generation

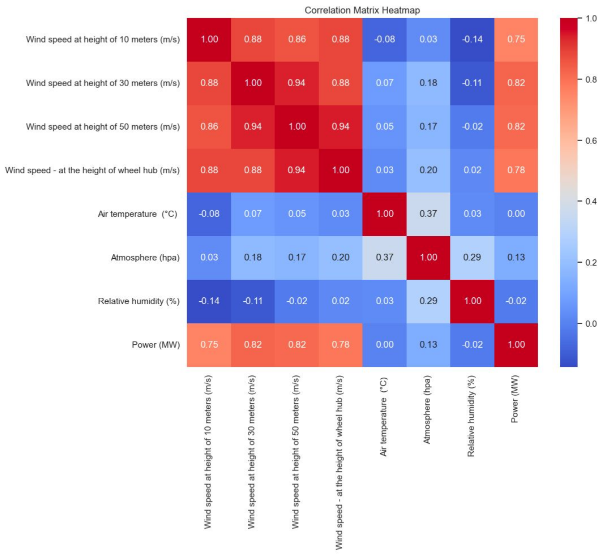

Fig. 4. Correlation matrix heatmap of wind data

VI. Conclusion

The journey through this study has provided valuable insights into the intricacies of predicting renewable energy production using machine learning models. While linear regression models served as a foundational framework, it became evident that more sophisticated techniques are required to fully capture the complexities inherent in renewable energy systems. The challenges encountered, particularly in the realm of nonlinear relationships and environmental variability, underscore the need for innovative approaches to enhance predictive accuracy.

The findings of this study serve as a roadmap for future research endeavors in renewable energy forecasting. By highlighting the strengths and limitations of various modeling techniques, this study lays the groundwork for further refinement and optimization of predictive algorithms. It is imperative to continue exploring advanced analytics methodologies, such as ensemble methods and deep learning architectures, to unlock the full potential of renewable energy forecasting.

Moreover, the implications of this research extend beyond academic curiosity to real-world applications. The integration of accurate renewable energy forecasts into energy management systems holds the key to achieving a sustainable and resilient energy future. By enabling more informed decision-making and resource allocation, advanced analytics empower stakeholders to navigate the complexities of energy transition with confidence and foresight.

In conclusion, this study reaffirms the pivotal role of advanced analytics in the optimization of renewable energy production. As we continue to innovate and refine our predictive models, we move closer to realizing the vision of a cleaner, greener, and more sustainable energy landscape for future generations.

VII. References

1) Yongbao Chen & Junjie Xu. Solar and wind power data from the Chinese State Grid Renewable Energy Generation Forecasting Competition. https://www.nature.com/articles/s41597-022-01696-6
2) Ceci, M., Corizzo, R., Fumarola, F., Ianni, M., Malerba, D., Maria, G., ... & Rashkovska, A. (2015, July). Big data techniques for supporting accurate predictions of energy production from renewable sources. In Proceedings of the 19th international database engineering & applications symposium (pp. 62-71).
3) Mujeeb, S., Alghamdi, T. A., Ullah, S., Fatima, A., Javaid, N., & Saba, T. (2019). Exploiting deep learning for wind power forecasting based on big data analytics. Applied Sciences, 9(20), 4417.
4) Buturache, A., & Stancu, S. (2021). Wind Energy Prediction Using Machine Learning. Low Carbon Economy. https://doi.org/10.4236/LCE.2021.121001.
5) Haupt, S., & Kosović, B. (2017). Variable Generation Power Forecasting as a Big Data Problem. IEEE Transactions on Sustainable Energy, 8, 725-732. https://doi.org/10.1109/TSTE.2016.2604679.
6) Je, S., Ko, H., & Huh, J. (2021). Accurate Demand Forecasting: A Flexible and Balanced Electric Power Production Big Data Virtualization Based on Photovoltaic Power Plant. Energies. https://doi.org/10.3390/en14216915.