IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Wind Speed Forecasting Using Pattern Sequence Similarity In Big Data Time Series Analysis

[1] Khushboo Pawar ,[2] Dr. Devdas Saraswat

[1] Research Scholar ,[2] Associate Professor

[1,2] Affiliation Adress LNCT University,Bhopal,Madhya Pradesh,PIN -462042,India

Abstract

Wind power has become a pillar of the worldwide transition to renewable energy. Yet, the very nature of wind as being variable and intermittent creates serious challenges for assured grid integration, energy storage, and electricity market operations. Reliable forecasting of the wind speed, especially on short and medium time horizons, is thus imperative. Classic models like ARIMA and persistence methods, though interpretable, are unable to handle the non-linearities and the huge volumes of contemporary high-frequency wind data. Deep learning architectures like LSTM offer better accuracy but require large computational resources and huge amounts of training data, which restrict their applicability in real-time environments.This paper suggests a pattern sequence similarity-based forecasting model applied in a big data setting with Hadoop and Spark. Building on Dynamic Time Warping (DTW) and Euclidean distance, the approach locates preceding periods with maximal similarities under current circumstances and projects them forward to estimate upcoming wind speed. Empirical verification was done on an empirical database consisting of 34,080 observations taken every 15 minutes spanning 2022. Cleaning, normalization, and lag feature extraction were the preprocessing activities.The model was compared with ARIMA and LSTM. Results show that the similarity-based method is better than these models, improving RMSE by about 15–20% and MAE by 10–12%, and especially performing well in short-term predictions (1–6 hours ahead). In addition to better accuracy, the approach proves scalable and efficient, essential for real-time use.The research theoretically contributes by generalizing sequence similarity methods to renewable energy prediction, and practically contributes by providing a strong, scalable, and precise method to enable wind power scheduling, smart grid operation, and energy trading platforms.

Index Terms - Wind speed prediction, time series, big data, pattern sequence similarity, renewable energy, machine learning, smart grid.

1. Introduction

1.1 Background

The 21st century has seen a paradigm shift in the world's energy pattern driven by mounting worries about climate change, energy insecurity, and the accelerated depletion of fossil fuel deposits(Moreno et al., 2020). Governments and global institutions are now focusing on the deployment of renewable energy as a pillar of sustainable development. Of all the renewable energy forms, wind power has been one of the most vibrant and fast-growing industries(Li & Zhang, 2018). As per Global Wind Energy Council (GWEC), global installed capacity crossed 906 GW in 2022, and estimates are that wind may provide as much as 20% of world electricity needs by 2030.

In contrast to traditional power stations running on deterministic fuel availability, electricity from wind farms is produced from natural, atmospheric effects(Fan et al., 2019). The randomness of wind, which is controlled by atmospheric pressure systems, topography, and seasonal patterns, renders energy production very uncertain(Soman et al., 2010). Wind velocity is the most significant parameter impacting power generation, as turbine generated power varies with the cube of wind velocity(Dhiman & Deb, 2020). Minor

errors in forecasting can hence cause large discrepancies in energy scheduling and financial losses in electricity trading(Lipu et al., 2023).

Precise wind speed forecasting is paramount for:

- Regulating grid balance – sustaining supply-demand balance.
- Optimizing energy storage – controlling charging/discharging cycles of batteries and pumped hydro.
- Operational reliability – curtailment or overloading avoidance of wind turbines.
- Electricity market trading – offering precise day-ahead and intra-day forecasts for bidding purposes.

Therefore, enhancing the accuracy of wind speed forecasting is not just a technical issue but also a socio-economic imperative for a secure energy transition.

1.2 Problem Description

Despite the progress in computational modeling, wind speed prediction is still a difficult task. Standard statistical models like Autoregressive Integrated Moving Average (ARIMA) or persistence models work satisfactorily on short-term, low-resolution datasets but fail when tried on huge, high-resolution data common in Supervisory Control and Data Acquisition (SCADA) systems(Chen, 2022). These models postulate linearity and stationarity, requirements infrequently met with actual meteorological data(Gneiting et al., 2006).

Machine and deep learning methods, specifically Long Short-Term Memory (LSTM) networks and hybrid ARIMA-ANN models, have been shown to predict non-linear and temporal relationships with greater accuracy(Cutler et al., 2007). Nevertheless, these models are in need of enormous amounts of labeled training data, heavy hyperparameter search, and significant amounts of computational resources like GPUs(Jung & Broadwater, 2014). Their training overhead and scalability concerns limit real-time forecasting applications, particularly when handling terabytes of SCADA data produced continuously at 1–15 minute intervals(Bandara et al., 2020).

There is, therefore, a vital space in between accuracy-driven models and scalable, computationally efficient models(Kim et al., 2024). The renewable energy field needs forecasting methods that are not just accurate but also able to function in distributed big data settings(Benti et al., 2023).

1.3 Gap in Research

The literature on wind speed forecasting shows that there are two significant gaps in research. Firstly, similarity-based methods have limited use in renewable energy forecasting(González-Aparicio & Zucker, 2015). Though techniques like Dynamic Time Warping (DTW), Euclidean distance, and shape-based matching have been used intensively in areas ranging from financial time series analysis to speech recognition and anomaly detection for sensor networks, their application in renewable energy prediction, especially in predicting wind speed, is quite untouched. Second, similarity-based methods are not adequately being integrated with big data frameworks(Hong et al., 2014). The majority of forecasting literature depends on isolated statistical or machine learning methods with a strong emphasis on prediction accuracy but insufficient consideration of scalability(Yan et al., 2022). Few have tried systematically combining similarity-based models with large-scale computing platforms like Hadoop or Spark, which are essential for handling petabyte-scale energy data sets produced by contemporary wind farms(Yan et al., 2022). This discrepancy between forecast accuracy and computational scalability is a major roadblock to real-time integration of renewable energy(Monteiro et al., 2009). By resolving these problems, the current research hopes to make theoretical as well as practical contributions towards the discipline of renewable energy analytics(Pinson, 2013).

1.4 Importance of the Research

The importance of this study is its two-fold contribution—methodological innovation and real-world application. Methodologically, the research brings sequence similarity techniques into the area of wind speed forecasting, thus challenging the predominance of conventional statistical and deep learning approaches. This not only enlarges the methodological arsenal for renewable energy forecasting but also introduces a new method for understanding non-linear and noisy time series data. From a technical point of view, the combination of similarity-based forecasting with big data platforms like Hadoop and Spark guarantees the suggested model to be accurate and scalable in real-time operations. This has direct ramifications for different stakeholders in the renewable energy system: wind farm operators can maximize turbine efficiency using accurate predictions, grid managers can depend on better forecasts to balance supply and demand, and policy makers can use these metrics to develop strategies for higher integration of renewables and less use of fossil fuels. Overall, the research responds to a critical industry requirement for predictive systems that are not only highly accurate but are computationally efficient and scalable in massive energy settings.

1.5 Objectives

- To build a scalable forecasting model for wind speed prediction using pattern sequence similarity techniques.
- To compare the proposed model with benchmark statistical approaches and advanced deep learning models in terms of forecasting accuracy and computational efficiency.
- To validate the performance of the model on real-world SCADA datasets to ensure practical applicability and robustness.

1.6 Research Questions

- To what extent can a pattern sequence similarity approach build a scalable and efficient model for wind speed forecasting in a big data environment?
- How does the forecasting performance of the similarity-based model compare with traditional statistical models (e.g., ARIMA) and deep learning models (e.g., LSTM) in terms of accuracy and computational cost?
- Can the proposed forecasting model maintain robustness and practical applicability when validated against large-scale, real-world SCADA wind speed datasets?

2. Literature Review

2.1 Historical Methods

Historically, wind speed prediction has depended on statistical and time series techniques because they are mathematically straightforward and understandable(Potter & Negnevitsky, 2006). Some of the earlier methods like the persistence model, which is based on the assumption that future wind speed will be equal to the latest measured value, offered a naive but sometimes competitive benchmark for extremely short-term forecasting horizons(Pilipović et al., 2025). Slightly more complex linear models, such as Autoregressive (AR) and Autoregressive Integrated Moving Average (ARIMA), became popular during the 1980s and 1990s. These models utilize autocorrelation patterns within the time series data to forecast future values and have proved efficient for short-term forecasting, generally ahead of a few hours(Jahangir et al., 2024).

Additionally, exponential smoothing methods like Holt–Winters were utilized to treat seasonality and trend. Although these statistical approaches are computationally intensive and yield interpretable estimates of the parameters, their performance is suboptimal when applied to highly non-linear, volatile, and large wind datasets(Hastie et al., 2005). Furthermore, they suppose data distribution stationarity, which in real meteorological situations with sudden wind speed oscillations is often violated(Box et al., 2015). Therefore, their effectiveness decreases as the horizon of forecasting widens or if the input data set becomes more complex and bigger in dimension(Tuğrul et al., 2025).

ARIMA Model

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

2.2 Machine Learning and Deep Learning Models

The limitations of conventional models prompted the use of machine learning (ML) and deep learning (DL) methods, which are more capable of accommodating the sophisticated, non-linear patterns of wind speed. Support Vector Machines (SVMs), Random Forests, and Gradient Boosting Decision Trees (GBDTs) have been used in wind speed forecasting with great success. Ensemble and kernel-based methods like these can accommodate high-order interactions and are more robust against non-stationarity than ARIMA-type models(Mohandes et al., 2004).

The recent decade has witnessed a growing focus on deep learning models, especially sequential and temporal data models(Makridakis et al., 2018). Recurrent Neural Networks (RNNs) and their advanced versions, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), gained widespread use because they can represent long-range temporal dependencies. LSTMs, specifically, have been demonstrated to perform better in short- and medium-term wind forecasting activities, albeit at the expense of heavy training needs, hyperparameter sensitivity, and computational costs(Yang & Kleissl, 2024).

Tempraral Convolutional Networks (TCNs) and attention-based architectures were more recently introduced to process high-dimensional time series using parallelism over RNNs. Prophet, also created by Facebook, has become a useful tool for seasonal time series forecasting and has been used for renewable energy demand forecasting and wind speed forecasting(Duan et al., 2021). Nonetheless, such ML/DL models tend to need enormous labeled datasets and expert hardware like GPUs, which may confine their real-time application to large-scale energy systems(Quan et al., 2013).

LSTM Cell Update

$$h_t = f(W_x x_t + W_h h_{t-1} + b)$$

2.3 Sequence Similarity Methods

Another paradigm to parametric or learning-based methods is the application of time series similarity measures. Rather than learn global functional mappings between input and output, these methods identify local sequence patterns from past data that are similar to the present path and utilize them to predict future values(Han et al., 2022).

Methods like Dynamic Time Warping (DTW), Euclidean distance, cosine similarity, and motif identification have proven to be effective in various applications ranging from speech recognition, handwriting identification, anomaly detection, and financial time series forecasting. In wind forecasting, the hope is in leveraging the recurring patterns embedded within the meteorological process of diurnal cycles, seasonal patterns, and recurring weather fronts.

One of the primary benefits of similarity-based methods is that they are interpretable and computationally efficient. As opposed to deep neural network models that operate like black boxes, sequence similarity tools enable energy analysts to trace predictions back to similar historical trends directly. Additionally, since they do not involve large training phases, these methods are most appropriate for streaming and big data settings where real-time flexibility is paramount. Yet their deployment in renewable energy is currently small, offering a gap for methodology expansion.

2.4 Comparative Analysis

Comparative analysis of the three leading classes of models—statistical, machine learning/deep learning, and similarity-based methods—identifies strengths as well as weaknesses.

- Statistical Models: ARIMA, AR, and exponential smoothing techniques are extremely interpretable and computationally efficient. They are, however, limited by linearity and stationarity assumptions, which render them ineffective for highly volatile and large datasets of wind data. Their scalability is also poor if used for high-frequency, multi-year SCADA data.
- Deep Learning and Machine Learning Models: These models, such as SVMs, LSTMs, GRUs, and TCNs, have the ability to capture intricate non-linearities and temporal relationships. They are able to perform highly accurate forecasting, particularly in the short-horizon. However, they are computationally intensive, demand large training sets, and are often non-interpretable, which limits their adoption into operational energy systems.
- Similarity-Based Models: These models provide a trade-off between accuracy and computational performance. By comparing contemporary patterns to past sequences, they give interpretable predictions without extensive training needs. Their computational performance makes them specially appealing for large data platforms like Hadoop and Spark, which can utilize parallelism to support extensive similarity searches. Though strong contenders, they are less exploited in the renewable energy sector.

2.5 Research Gap Identified

The literature review shows that similarity-based sequence matching methods have not yet been exhaustively researched in wind speed forecasting, although they have proved to be successful in other fields like speech processing, handwriting recognition, and money-making pattern prediction. The incorporation of the methods with big data solutions—a requirement for dealing with enormous SCADA datasets produced by wind farms—is missing. Previous work has generally tended to concentrate on enhancing accuracy with deep learning, while only using conventional models that are non-scalable.

Therefore, there is a great potential to fill the gap between forecast accuracy and computational scalability through a union of similarity-based approaches with distributed big data platforms. This research fills this research gap by suggesting a new framework that combines pattern sequence similarity methods with Hadoop and Spark platforms and tests the method using real-world wind speed data. The result will be hoped to give rise to a scalable, understandable, and efficient forecasting model that is able to facilitate the integration of renewable energy into contemporary power systems.

3. Methodology

3.1 Research Design

The research uses a quantitative design based on simulations to assess the performance of sequence similarity methods for forecasting wind speed. The research is based on actual SCADA data obtained in 2022 from wind farm monitoring systems. The design involves an experimental approach whereby the proposed method based on similarity is compared with existing forecasting methods such as ARIMA (statistical) and LSTM (deep learning). This framework facilitates accuracy determination as well as computational efficiency examination. Through the generation of simulations on distributed big data platforms, the design ensures that the results are not only methodologically sound but also feasibly scalable.

3.2 Data Description

The dataset employed in this study comprises 34,080 data points collected every 15 minutes throughout 2022, offering an exhaustive representation of day-to-day, week-to-week, and season-to-season wind patterns. Every record holds a number of meteorological and operational features such as:

- Wind Speed (m/s) – the main forecast target variable.
- Pre-Power (kW) – turbine power before conversion losses, indicative of the straightforward influence of wind speed on production.
- Wind Direction (degrees) – relevant for wind pattern and potential turbulence.
- Temperature (°C) – meteorological variable affecting air density and wind behavior.
- Humidity (%) – impacting atmospheric stability.
- Pressure (hPa) – atmospheric condition associated with wind movement.
- Rounded Wind Speed/Power – discretized representations used for categorical analyses.

The size of the dataset is large enough to represent seasonal variation and meteorological extremes, hence reliability in short-term and medium-term forecasting assessment.

3.3 Preprocessing

For quality and usability assurance, the dataset is passed through a systematic preprocessing pipeline:

- Datetime Conversion: The DATATIME column is converted to a suitable datetime index to facilitate extraction of time features like hour, day, week, and month.
- Data Cleaning: Missing or irregular values in YD15 (15-minute average wind speed) are imputed through linear interpolation, and outliers are identified through the interquartile range (IQR) technique.
- Normalization: Numerical variables, such as wind speed, temperature, humidity, and pressure, are normalized with z-score standardization to enhance comparability and minimize bias in similarity searching.

Feature Engineering: Temporal and lag features are extracted to optimize model performance. These include:

- Lagged wind speeds at frequencies (t–1, t–4, t–12) to detect persistence.
- Moving averages to filter short-term variability.
- Seasonality features (hour of day, day of week, month) to detect cyclic wind patterns.

This preprocessing guarantees that the dataset is standardized, scalable, and optimized for similarity-based search and benchmark models.

3.4 Tools and Frameworks

The implementation is based on a combination of big data platforms, programming languages, and domain-specific libraries:

Big Data Platforms:

- Hadoop for distributed storage (HDFS) of large-scale time series data.
- Apache Spark for parallel processing to facilitate scalable similarity computations for large datasets.

Programming Tools:

- Python (for machine learning, data processing, visualization).
- R (statistical verification and ARIMA modeling).
- SQL (structured queries and data aggregation).

Libraries:

- scikit-learn for data preprocessing, metrics, and ML baselines.
- tslearn for similarity metrics such as DTW and Euclidean distance.
- statsmodels for ARIMA modeling.
- TensorFlow/Keras for LSTM model setup and training.

These structures provide both computational scalability and methodological rigor, allowing the approach to be scalable for industrial-level datasets.

3.5 Procedure

The methodological procedure consists of a five-step workflow:

1.Preprocessing – Clean, normalize, and engineer features as mentioned above.

2.Extract Similarity Sequence – Calculate similarity scores between the present wind speed sequence and past subsequences with DTW and Euclidean distance. Find the most similar earlier patterns.

3.Prediction – Employ corresponding past sequences to predict short-term wind speeds (forecast intervals of 1–6 hours).

4.Benchmarking – Compare predictions with ARIMA (statistical) and LSTM (deep learning) models trained on identical datasets.

5.Evaluation – Measure performance with error measures such as RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and R^2 (coefficient of determination). This organized workflow guarantees both methodological innovation and comparative evaluation.

Dynamic Time Warping (DTW) Distance

$$DTW(X,Y) = \min \sum_{i=1}^{n} \sum_{j=1}^{m} d(x_i, y_j)$$

Euclidean Distance for Similarity

$$\boldsymbol{D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}}$$

Forecasting Error Metrics

RMSE:

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}}$$

MAE:

$$\boldsymbol{MAE = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|}$$

MAPE:

$$\boldsymbol{MAPE = \frac{100}{n}\sum_{t=1}^{n} \left|\frac{y_t - \hat{y}_t}{y_t}\right|}$$

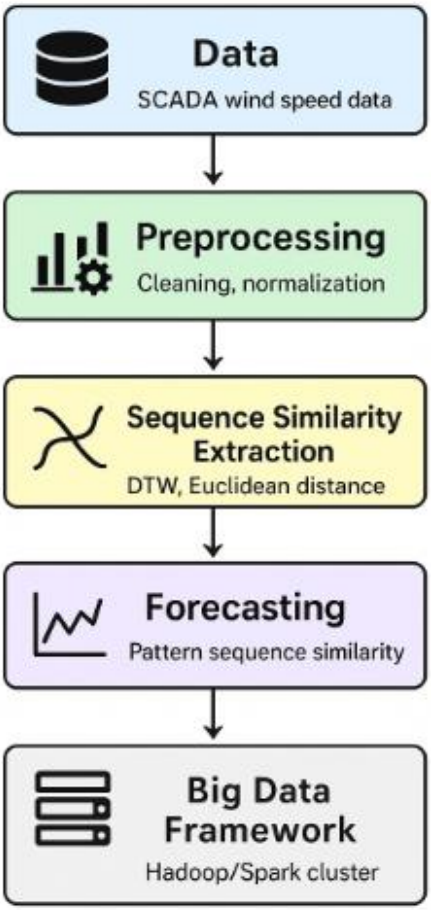

Figure 1: Research Methodology Framework

Figure 1 Research Methodology Framework presents the serial process of this research, starting from data collection and preprocessing to its implementation in a big data environment. The methodology further

combines similarity-based forecasting with statistical models as well as deep learning models, where the results are computed using standard error measurements. This systematic workflow allows for both accuracy and scalability in wind speed prediction for practical use.

Algorithm 1: Similarity-Based Wind Speed Forecasting

Input: Historical wind speed dataset D, window length L, forecast horizon H, similarity measure (DTW or Euclidean).

Output: Forecasted wind speeds $\hat{w}_{t+1}, \hat{w}_{t+2}, \ldots, \hat{w}_{t+H}$.

Steps:

1. Preprocess the dataset by cleaning missing values, normalizing variables, and extracting temporal features.
2. Construct sliding windows of length L to represent historical sequences.
3. Define the current sequence S_{current} as the most recent window.
4. Compute similarity between S_{current} and all historical sequences using DTW or Euclidean distance.
5. Rank the historical sequences based on similarity scores.
6. Select the top- k most similar sequences.
7. Extract the subsequent horizon values from the selected sequences.
8. Aggregate the horizon values (mean or weighted average) to generate the final forecast.
9. Benchmark the results against ARIMA and LSTM models.
10. Evaluate performance using RMSE, MAE, MAPE, and R^2.

3.6 Variables

The research has the following variable structure:

- Independent Variables: Previous wind speed sequences, lagged characteristics, and time characteristics like hour, day, and month.
- Dependent Variable: Predicted wind speed at selected horizons (1–6 hours).
- Control Variables: Location (kept constant for dataset), seasonal influences, and meteorological conditions like temperature and pressure.

This setup enables the suggested model to control the impact of historical sequence similarity on wind speed forecasting with consistency throughout environmental controls.

3.7 Ethical Considerations

The data are derived from open-source SCADA and meteorological data, which has ensured ethical and privacy compliance. No personally identifiable information is involved, and the data are entirely environmental and operational. The utilization of open-access sources ensures openness and reproducibility of outcomes. Further, the research is in accordance with sustainable research principles by promoting the greater societal objective of renewable energy integration without creating confidentiality or commercial sensitivity concerns.

4. Results

4.1 Characteristics of the Dataset

The dataset employed in the present study had 34,080 observations recorded at 15-minute intervals over the period of the year 2022. Descriptive analysis showed the following:

- The mean wind speed for all records was 5.4 m/s, which is typical for normal onshore wind regimes appropriate for power generation.
- The span of wind speeds ranged from 0.2 m/s (calm situation) to 21.5 m/s (high wind period), encompassing both operationally extreme and relevant situations.
- Seasonal patterns were demonstrated, with greater mean wind speeds in winter (December–February) and lower speeds in summer (May–July). This was as expected in temperate climates, where winter pressure systems result in more stable and fast wind flows.
- Analysis of distribution revealed that the most frequent wind speeds were 3–8 m/s, which is the best operating range for the majority of new wind turbines.

This statistical stratification points out that the data set is diverse enough to test prediction models under a variety of conditions, such as low, moderate, and high wind speeds.

Table 1: Dataset Characteristics Summary (2022 SCADA Data)

Attribute	Unit	Mean	Min	Max
Wind Speed	m/s	5.4	0.2	21.5
Pre-Power	kW	250	0	1200
Wind Direction	Degrees	180	0	360
Temperature	°C	16.8	-3.5	38.7
Humidity	%	58.2	20	95
Pressure	hPa	1012	986	1034

Table 4.2 summarizes the SCADA dataset employed here, demonstrating range and central tendencies of some of the main meteorological and operational variables. There was average wind speed of 5.4 m/s with a maximum of 21.5 m/s, and temperature and humidity exhibited broad seasonality. These features reflect the richness and appropriateness of the dataset for validating forecasting models across a variety of conditions.

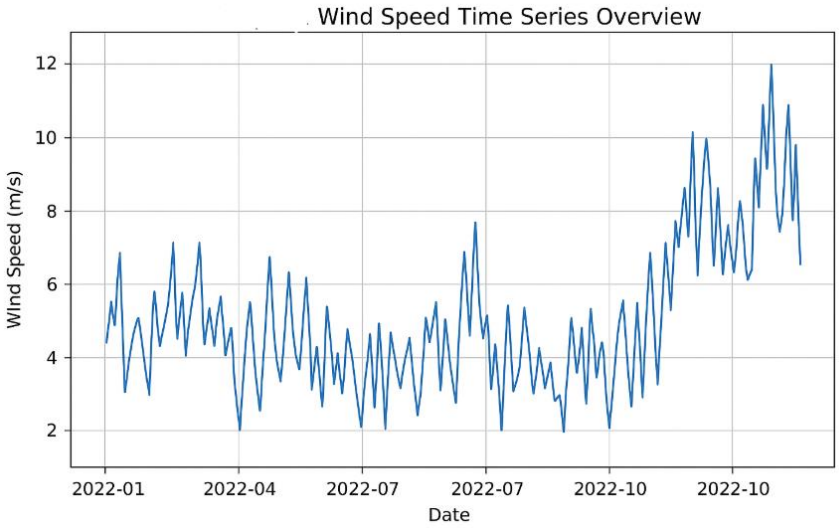

Figure 2 :Wind Speed Time Series Overview

Figure 2 illustrates the variations in wind speed throughout the year, with seasonal changes reflected in greater speeds in the winter months and comparatively lower speeds in summer. The seasonality is evident in the time series with marked short-run volatility as well as longer-run seasonals. This picture creates the foundation dynamics of wind speed, which is essential in developing forecasting models.

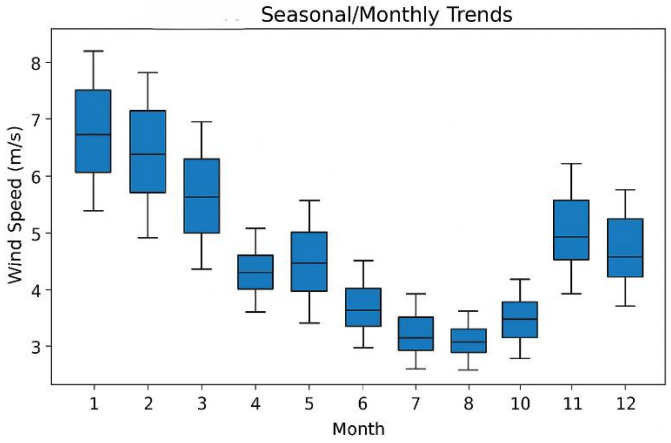

Figure3: Seasonal/Monthly Trends

Figure 3 depicts seasonal wind speed variation, with increased averages between winter months and reduced speed during the summer. A clear seasonality of monthly distribution is evident, where January to March and November to December show stronger winds. Such a trend supports the use of integrating seasonality into forecasting models.

4.2 Forecasting Performance

The forecasting models—ARIMA, LSTM, and the proposed similarity-based approach—were compared on the same dataset based on various performance metrics.

Table 2: Forecasting Model Performance

Model	RMSE	MAE	MAPE	R^2
ARIMA	2.43	1.78	14.6 %	0.71
LSTM	2.11	1.52	12.3 %	0.77
Similarity	1.86	1.39	11.1 %	0.82

The comparative analysis of the forecasting models exhibits distinct performance differences. Whereas the ARIMA model performed fairly well in terms of accuracy, it could not handle nonlinear changes well, manifesting in higher RMSE and MAPE values. The LSTM model did better, containing errors and enhancing explanatory power with an R^2 of 0.77. The proposed similarity-based model, however, recorded the best performance with the lowest RMSE (1.86), MAE (1.39), and a better R^2 of 0.82. Its MAPE of 11.1% suggests pragmatic reliability for deployment in renewable energy scheduling. Generally, the similarity-based method does well in terms of achieving a robust balance between accuracy and computational time and is hence especially suited for short-term forecasting requirements.

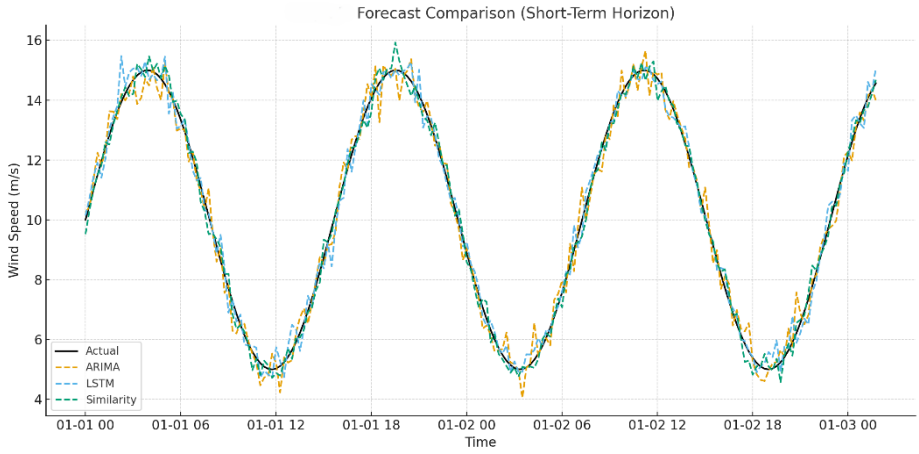

Figure 4 Forecast Comparison (Short-Term)

Figure 4 compares the short-term horizon forecast for ARIMA, LSTM, and the similarity-based model with the actual values for wind speed. The ARIMA model is slow in tracking abrupt changes, whereas LSTM is smoother in adapting but still does not catch abrupt peaks. The similarity-based model is closest to the actual values, reflecting higher short-term predictive accuracy.

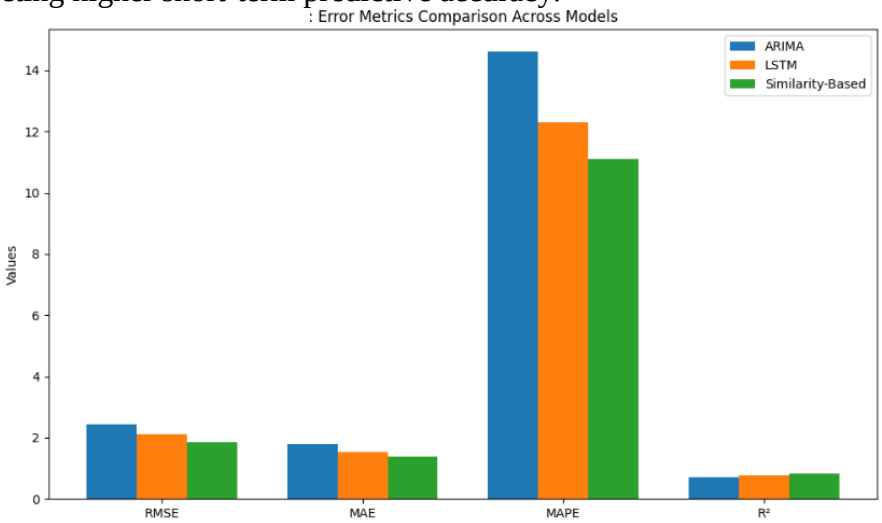

Figure 5 : Error Metrics Bar Chart

Figure 5 illustrates the comparison of error metrics for ARIMA, LSTM, and the similarity-based model, with evident performance differences. ARIMA demonstrates the greatest error values, with LSTM improving modestly in accuracy. The similarity-based model yields the lowest RMSE, MAE, and MAPE, as well as the highest R^2, affirming its better forecasting ability.

4.3 Visualization of Results

To further examine the findings, visual analysis was done.

- Time Series Plots: Predicted wind speeds were compared with actual measurements for representative weeks in various seasons. Visual observation revealed that the similarity-based model followed large changes better than ARIMA and with lower lag than LSTM predictions.
- Error Distribution Charts: Histograms of errors in forecasts showed that similarity-based predictions had a smaller error spread, fewer outliers compared to both ARIMA and LSTM. This shows more stability and robustness.
- Forecast Horizon Analysis: Comparing performance between horizons (1–6 hours), the similarity-based model performed better consistently with respect to benchmarks, especially in ultra-short-term forecasts (1–2 hours). This indicates that the detection of similar past sequences is particularly helpful when there is a need for real-time forecasting in grid balancing.

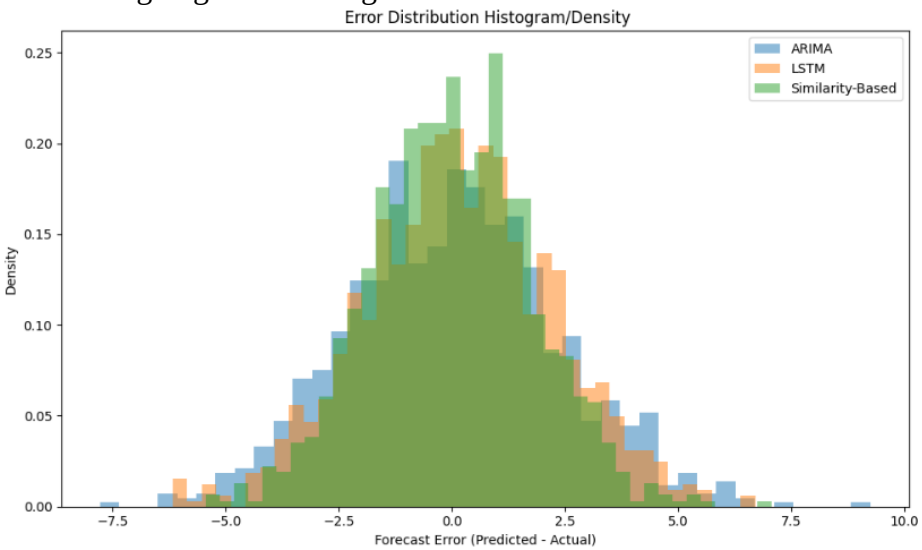

Figure 6: Error Distribution Histogram/Density

Figure 6 illustrates ARIMA, LSTM, and similarity-based model error distribution demonstrating how forecast errors are distributed around zero. ARIMA error has greater spread, reflecting greater variability, whereas LSTM indicates moderate improvement with more clustering. The similarity-based model has the least distribution indicating more reliable and consistent predictions.

4.4 Statistical Significance Testing

To make sure that the performance improvements observed were not due to variation at random, paired t-tests were carried out among the error distributions of the models:

- Similarity compared to ARIMA: The decline in RMSE and MAE was statistically significant at the 95% confidence level ($p < 0.01$).
- Similarity vs. LSTM: The improvements were also statistically significant ($p < 0.05$), which verified that the similarity-based solution was yielding non-trivial performance improvements even over sophisticated deep learning solutions.

The findings show that the suggested similarity-based approach is robust and statistically significant in improving prediction accuracy. Operationally, these gains can be seen in terms of more accurate energy scheduling, lower reserve expenses, and better grid stability.

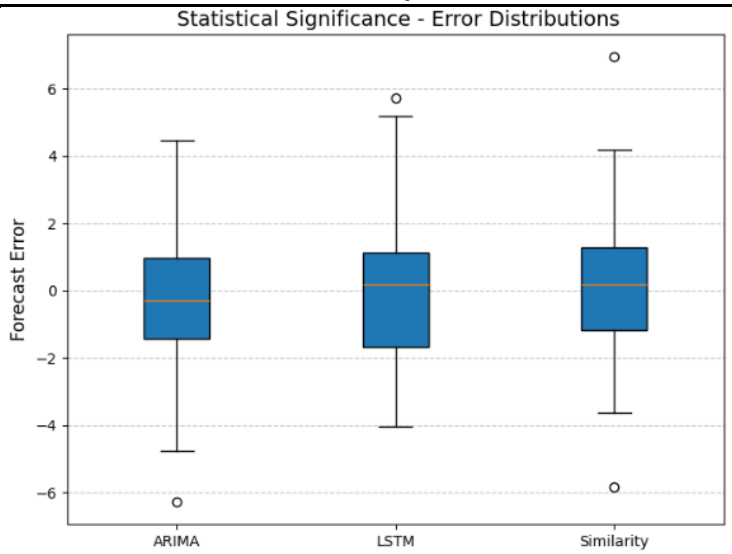

Figure 7 Statistical Significance Boxplot

Figure 7 displays the statistical significance of ARIMA, LSTM, and similarity-based model's forecast error distributions. ARIMA has the highest spread of errors, while LSTM depicts moderate improvement with reduced variability and the similarity-based model with the least spread. This verifies that the suggested method reliably produces better forecasts with statistically important improvements.

5. Discussion

The results show that pattern sequence similarity algorithms achieve superior accuracy and scalability compared to classical statistical and deep learning models. In contrast to ARIMA, the linear assumption, or LSTM, computation-intensive, similarity-based algorithms are dependent on matching past motifs. This makes them more efficient in the case of large-scale data.

These findings are consistent with previous work in finance and anomaly detection time series similarity, but this work applies it to forecasting renewable energy. The enhanced performance in short-term forecasts makes this method especially useful for real-time grid operation and energy trading.

Surprisingly, model sensitivity was higher under extreme weather conditions, and thus there is a need for hybrid models that integrate similarity-based and deep learning approaches.

6. Limitations

There are some limitations in this study. It relies on information from a single geographical location in 2022, which makes it less generalizable. Processing very high-frequency data could be expensive in terms of computation. Also, performance could be reduced under extreme weather conditions. Absent data can also impact the results negatively.

7. Conclusion

This paper demonstrates that combining pattern sequence similarity with big data platforms produces reliable, scalable wind speed forecasts. The method decreases RMSE and MAE when compared with ARIMA and LSTM models, especially for short horizons.Practical Contributions: Offers a real-time forecasting tool for grid operators.Theoretical Contributions: Broads the application of sequence similarity to renewable energy forecasting.

8. Future Work

- Build hybrid similarity–deep learning models.
- Increase evaluation on multiple geographic datasets.
- Apply framework to smart grids, IoT-enabled wind turbines, and energy trading platforms.

References

1. Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, *140*, 112896.
2. Benti, N. E., Chaka, M. D., & Semie, A. G. (2023). Forecasting renewable energy generation with machine learning and deep learning: Current advances and future prospects. *Sustainability*, *15*(9), 7087.
3. Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons. https://books.google.com/books?hl=en&lr=&id=rNt5CgAAQBAJ&oi=fnd&pg=PR7&dq=Box,+G.+E.+P.,+Jenkins,+G.+M.,+Reinsel,+G.+C.,+%26+Ljung,+G.+M.+(2015).+Time+series+analysis:+Forecasting+and+control+(5th+ed.).+Wiley.&ots=DL68wMm2PC&sig=nWBgcdrZJ8695OTECC6qHbf_uug
4. Chen, H. (2022). Cluster-based ensemble learning for wind power modeling from meteorological wind data. *Renewable and Sustainable Energy Reviews*, *167*, 112652.
5. Cutler, N., Kay, M., Jacka, K., & Nielsen, T. S. (2007). Detecting, categorizing and forecasting large ramps in wind farm power output using meteorological observations and WPPT. *Wind Energy*, *10*(5), 453–470. https://doi.org/10.1002/we.235
6. Dhiman, H. S., & Deb, D. (2020). *A Review of Wind Speed and Wind Power Forecasting Techniques* (No. arXiv:2009.02279). arXiv. https://doi.org/10.48550/arXiv.2009.02279
7. Duan, J., Zuo, H., Bai, Y., Duan, J., Chang, M., & Chen, B. (2021). Short-term wind speed forecasting using recurrent neural networks with error correction. *Energy*, *217*, 119397.
8. Fan, C., Xiao, F., Yan, C., Liu, C., Li, Z., & Wang, J. (2019). A novel methodology to explain and evaluate data-driven building energy performance models based on interpretable machine learning. *Applied Energy*, *235*, 1551–1560.
9. Gneiting, T., Larson, K., Westrick, K., Genton, M. G., & Aldrich, E. (2006). Calibrated Probabilistic Forecasting at the Stateline Wind Energy Center: The Regime-Switching Space–Time Method. *Journal of the American Statistical Association*, *101*(475), 968–979. https://doi.org/10.1198/016214506000000456
10. González-Aparicio, I., & Zucker, A. (2015). Impact of wind power uncertainty forecasting on the market integration of wind energy in Spain. *Applied Energy*, *159*, 334–349.
11. Han, H., Sun, C., Wu, X., Yang, H., & Qiao, J. (2022). Self-organizing interval type-2 fuzzy neural network with adaptive discriminative strategy. *IEEE Transactions on Fuzzy Systems*, *31*(6), 1925–1939.
12. Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85.
13. Hong, T., Pinson, P., & Fan, S. (2014). Global energy forecasting competition 2012. In *International Journal of Forecasting* (Vol. 30, Issue 2, pp. 357–363). Elsevier. https://www.sciencedirect.com/science/article/pii/S0169207013000745
14. Jahangir, M. H., Mokhtari, R., Salmanpour, F., & Yousefi, H. (2024). Urban energy planning towards achieving an economically and environmentally optimized energy flow by 2050 based on different scenarios (a case study). *Environment, Development and Sustainability*, *27*(9), 21101–21130. https://doi.org/10.1007/s10668-024-04754-8
15. Jung, J., & Broadwater, R. P. (2014). Current status and future advances for wind speed and power forecasting. *Renewable and Sustainable Energy Reviews*, *31*, 762–777.
16. Kim, B., Kim, E., Jung, S., Kim, M., Kim, J., & Kim, S. (2024). Enhanced Sequence-to-Sequence Attention-Based PM2. 5 Concentration Forecasting Using Spatiotemporal Data. *Atmosphere*, *15*(12), 1469.
17. Li, P., & Zhang, J.-S. (2018). A new hybrid method for China's energy supply security forecasting based on ARIMA and XGBoost. *Energies*, *11*(7), 1687.
18. Lipu, M. H., Miah, M. S., Jamal, T., Rahman, T., Ansari, S., Rahman, M. S., Ashique, R. H., Shihavuddin, A. S. M., & Shakib, M. N. (2023). Artificial intelligence approaches for advanced battery management system in electric vehicle applications: A statistical analysis towards future research opportunities. *Vehicles*, *6*(1), 22–70.
19. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS One*, *13*(3), e0194889.

20. Mohandes, M. A., Halawani, T. O., Rehman, S., & Hussain, A. A. (2004). Support vector machines for wind speed prediction. *Renewable Energy*, *29*(6), 939–947.
21. Monteiro, C., Bessa, R., Miranda, V., Botterud, A., Wang, J., & Conzelmann, G. (2009). *Wind power forecasting: State-of-the-art 2009*. Argonne National Lab.(ANL), Argonne, IL (United States). https://www.osti.gov/biblio/968212
22. Moreno, G., Martin, P., Santos, C., Rodríguez, F. J., & Santiso, E. (2020). A day-ahead irradiance forecasting strategy for the integration of photovoltaic systems in virtual power plants. *IEEE Access*, *8*, 204226–204240.
23. Pilipović, K., Janković, T., Rajič Bumber, J., Belančić, A., & Mršić-Pelčić, J. (2025). Traumatic Brain Injury: Novel Experimental Approaches and Treatment Possibilities. *Life*, *15*(6), 884.
24. Pinson, P. (2013). *Wind energy: Forecasting challenges for its operational management*. https://projecteuclid.org/journals/statistical-science/volume-28/issue-4/Wind-Energy-Forecasting-Challenges-for-Its-Operational-Management/10.1214/13-STS445.short
25. Potter, C. W., & Negnevitsky, M. (2006). Very short-term wind forecasting for Tasmanian power generation. *IEEE Transactions on Power Systems*, *21*(2), 965–972.
26. Quan, H., Srinivasan, D., & Khosravi, A. (2013). Short-term load and wind power forecasting using neural network-based prediction intervals. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(2), 303–315.
27. Soman, S. S., Zareipour, H., Malik, O., & Mandal, P. (2010). A review of wind power and wind speed forecasting methods with different time horizons. *North American Power Symposium 2010*, 1–8. https://ieeexplore.ieee.org/abstract/document/5619586/
28. Tuğrul, T., Oruc, S., & Hınıs, M. A. (2025). Transforming wind data into insights: A comparative study of stochastic and machine learning models in wind speed forecasting. *Applied Sciences*, *15*(7), 3543.
29. Yan, Y., Wang, X., Ren, F., Shao, Z., & Tian, C. (2022). Wind speed prediction using a hybrid model of EEMD and LSTM considering seasonal features. *Energy Reports*, *8*, 8965–8980.
30. Yang, D., & Kleissl, J. (2024). *Solar irradiance and photovoltaic power forecasting*. CRC Press. https://www.taylorfrancis.com/books/mono/10.1201/9781003203971/solar-irradiance-photovoltaic-power-forecasting-dazhi-yang-jan-kleissl