

A Leakage-Aware AutoML Framework for Wind Power Forecasting: Assessing Physical Consistency via SHAP

Tuğçe İnağ

Department of Industrial Engineering
Gazi University, Ankara, Türkiye
tugceduzce@gazi.edu.tr

Büşra Bişkin

Department of Industrial Engineering
National Defense University, Ankara, Türkiye
busrabiskin@gmail.com

Gül Didem Batur Sir

Department of Industrial Engineering
Gazi University, Ankara, Türkiye
dbatur@gazi.edu.tr

Abstract

Wind power forecasting is important for renewable energy integration, grid stability, and operational planning. However, model performance may be overestimated when temporal data leakage occurs or when historical power values are used without careful validation. This study proposes a leakage-aware AutoML framework for wind power forecasting using only instantaneous wind speed and wind direction information. The publicly available Wind Turbine SCADA Dataset from Kaggle was used. The dataset was sorted chronologically and divided into training and test sets using an 80%–20% split. Outliers were replaced with median values using thresholds derived from the training set. Wind direction was transformed into sine and cosine components to represent its circular nature. AutoGluon was used to compare multiple regression models under the same experimental setting. The selected Weighted Ensemble model achieved an RMSE of 505.90 kW, MAE of 250.63 kW, and R^2 of 0.8579 on the chronological test set. For interpretability, SHAP analysis was conducted using the LightGBM model, which achieved a similar R^2 of 0.8560. SHAP results showed that wind speed was the dominant predictor, while wind direction components had secondary effects. The cubic fit of wind speed SHAP contributions achieved an R^2 of 0.9231, supporting the physical consistency of the learned nonlinear relationship. The results show that leakage-aware AutoML combined with SHAP can provide accurate and interpretable wind power forecasts.

Keywords

Wind Power Forecasting, AutoML, SHAP, Physical Consistency, Data Leakage

1. Introduction

Wind energy is one of the fastest-growing and most cost-effective renewable energy sources in the transition toward a sustainable energy future (Han et al. 2020; Loza et al. 2024). High-accuracy short-term power forecasting models are needed to safely integrate wind energy systems into the grid and to improve operational efficiency. In recent years, machine learning (ML) algorithms such as Random Forest, LightGBM, and Deep Learning have achieved significant

success by outperforming traditional methods in modeling the complex nonlinear relationships between wind speed and power generation (Demolli et al. 2019).

However, the “black-box” nature of these artificial intelligence models with high prediction accuracy limits the trust of field operators and stakeholders in these technologies by reducing the transparency and reliability of the systems (Liao et al. 2024). To overcome this trust problem, Explainable Artificial Intelligence (XAI) methods, especially SHAP (SHapley Additive exPlanations), have been increasingly preferred to explain the logic behind predictions and to verify whether the model learns in accordance with physical rules (Lundberg and Lee 2017; Cakiroglu et al. 2023).

Two main methodological limitations are observed in the existing wind power forecasting literature. The first is the problem of “temporal data leakage,” which occurs when the chronological order of data is ignored and random splitting is used in the training of time series models (Bergmeir and Benítez 2012; Cerqueira et al. 2020). The second is the tendency of models to show autoregressive memorization by becoming overly dependent on historical production values (lagged variables), rather than learning the instantaneous aerodynamic effects of wind.

To address these limitations, this study proposes a leakage-aware machine learning framework that uses only instantaneous wind speed and wind direction information. In the developed model, the theoretical power curve and historical power production data, namely lagged variables, are intentionally excluded from the model. Thus, it is tested whether the algorithms can learn the relationship between instantaneous wind conditions and generated power without relying on past production values. Data leakage is reduced by preserving the timestamp of the dataset and splitting it chronologically into training and test sets. To determine the best prediction model, various algorithms are compared using an AutoML framework, and SHAP analysis is used to assess the physical consistency of the model.

1.1 Objectives

The main objective of this study is to develop a leakage-aware AutoML-based framework for wind power forecasting using only instantaneous wind speed and wind direction information. The specific objectives of the study are as follows:

- To design a chronological train-test evaluation strategy that prevents temporal data leakage in wind power forecasting.
- To evaluate the predictive performance of AutoML-based regression models using only instantaneous wind speed and wind direction, without using lagged power values or theoretical power curve information.
- To compare the performance of the selected AutoML ensemble model with an interpretable tree-based model suitable for SHAP analysis.
- To investigate whether the learned model behavior is physically consistent with the expected nonlinear relationship between wind speed and wind power generation through SHAP-based interpretation.

2. Literature Review

Wind power forecasting is a critical research area for renewable energy integration. In recent years, ML algorithms have increasingly replaced traditional forecasting methods in this field. Tree-based and kernel-based models such as Random Forest, LightGBM, and Support Vector Regression (SVR) have been found successful in modeling the complex nonlinear relationships between wind speed and generated power (Demolli et al. 2019; Alkesaiberi et al. 2022). In addition, deep learning (DL) techniques, particularly long short-term memory (LSTM) networks and convolutional neural networks (CNNs), have been widely used to capture sequential dependencies in time-series data (Ahmad and Zhang 2022). However, many high-accuracy forecasting models operate as “black-box” systems, raising important concerns for field operators in terms of operational reliability and transparency.

To address this transparency problem, researchers have increasingly adopted XAI methods, particularly SHAP. SHAP simplifies complex artificial intelligence decisions by clarifying the individual effects of input features on predictions and helps assess whether model outputs are consistent with domain knowledge (Lundberg and Lee 2017; Cakiroglu et al. 2024). Recent studies have shown that SHAP can be used to interpret wind forecasting models in relation to aerodynamic principles and can support operational safety tasks such as anomaly detection and feature stability analysis (Liao et al. 2024; Muhammed Anees et al. 2026).

Despite these advances, two major methodological gaps remain in the machine learning-based wind power forecasting literature. First, temporal data leakage may occur when the chronological order of time-series data is disrupted and random splitting is used during model training and testing (Bergmeir and Benitez 2012; Cerqueira et al. 2020). Second, many wind power forecasting models rely heavily on historical power production values or theoretical power curve information, leading to an autoregressive structure that may emphasize memorization rather than learning the instantaneous aerodynamic effects of wind.

In summary, although existing studies provide high prediction accuracy and XAI-based interpretability, there is still a need for leakage-aware forecasting frameworks that prevent copying from historical data through lag-free and chronological experimental designs. This study addresses this gap by using SHAP not only to explain model predictions, but also to quantitatively assess whether a model trained only with instantaneous wind conditions, namely wind speed and wind direction, can learn a physically consistent nonlinear relationship between wind speed and power generation.

3. Methods

This study proposes a leakage-aware machine learning framework for wind power prediction using only instantaneous meteorological inputs. The target variable is the measured active power output, while the input variables are limited to wind speed and wind direction. No lagged power values or historical delay variables are included in the model in order to evaluate whether the algorithms can learn the relationship between current wind conditions and power generation without relying on past output values.

Before model training, the dataset is sorted chronologically according to the timestamp. To reduce the risk of temporal data leakage, the data is divided into training and test sets using a chronological split, where the first 80% of observations are used for training and the remaining 20% are reserved for testing. Outlier treatment was performed after the chronological train-test split using only information derived from the training set. A wind-speed-bin-based approach was applied because the expected active power output depends strongly on wind speed. Wind speed was divided into 0.5 m/s intervals, and the median active power value was calculated separately for each wind speed bin in the training data. Within each bin, outlier thresholds were determined using the interquartile range (IQR) method. Observations with unusually high power values at low wind speeds were considered physically inconsistent and were replaced with the median active power value of their corresponding wind speed bin. In contrast, observations with low power values at high wind speeds were not removed, since such cases may represent realistic operational conditions such as turbine shutdowns, curtailment, maintenance, or faults. The bin-specific medians and IQR thresholds calculated from the training set were then applied to the test set when necessary, ensuring that no information from the test period was used during preprocessing.

Wind direction is transformed using sine and cosine functions to represent its circular nature. This transformation avoids the artificial discontinuity between 0° and 360°. Therefore, the final input variables consist of wind speed, sine-transformed wind direction, and cosine-transformed wind direction.

Several machine learning models were evaluated through the AutoGluon TabularPredictor AutoML framework. The target variable was defined as LV ActivePower (kW), and the problem type was set as regression. The evaluation metric was root mean squared error (RMSE). The AutoML process was run with the good_quality preset and a time limit of 1800 seconds. During training, AutoGluon evaluated several model families, including LightGBM, LightGBMXT, LightGBMLarge, XGBoost, Random Forest, Extra Trees, NeuralNetFastAI, NeuralNetTorch, and a level-2 Weighted Ensemble model. The final model was selected according to validation performance within the AutoGluon framework and was then evaluated on the unseen chronological test set. In addition to predictive accuracy, SHAP analysis was applied to interpret the contribution of each input variable to the model predictions. The final ensemble model selected by AutoGluon was a level-2 Weighted Ensemble that combined the predictions of LightGBMXT, LightGBM, XGBoost, and Extra Trees base learners. SHAP-based interpretability analysis was conducted using the LightGBM model, which achieved a very similar test performance and provided a suitable tree-based structure for feature contribution analysis.

4. Data Collection

The dataset used in this study was obtained from the publicly available Wind Turbine SCADA Dataset on Kaggle (Erisen, 2018). The dataset consists of wind turbine measurements recorded between January 1, 2018 and December

31, 2018. It includes 50,530 observations with a temporal resolution of 10 minutes. Each observation contains the timestamp, measured active power output, wind speed, theoretical power curve value, and wind direction.

The measured active power output, represented as LV ActivePower in kW, is selected as the target variable. Wind speed, measured in m/s, and wind direction, measured in degrees, are used as the input variables. Although the dataset also contains the theoretical power curve variable, this variable is excluded from the modeling process. This exclusion is intentional because the theoretical power curve already embeds prior information about the expected relationship between wind speed and power output. Including it as an input could make the prediction task artificially easier and weaken the evaluation of whether the model learns the physical relationship from the raw wind variables.

The original wind direction variable is transformed into sine and cosine components before modeling. This representation is more appropriate for directional data because wind direction is circular. For example, directions close to 0° and 360° are physically similar, although they appear numerically distant when represented as raw degrees.

The final modeling dataset therefore includes the measured active power output as the target variable and three input features: wind speed, sine of wind direction, and cosine of wind direction. The dataset is kept in chronological order throughout the analysis to preserve the temporal structure and to avoid data leakage between the training and test periods.

5. Results and Discussion

5.1 Numerical Results

The numerical performance of the proposed AutoML-based framework was evaluated on the chronological test set, corresponding to the final 20% of the dataset. Table 1 presents the test results of the best AutoGluon ensemble model and the LightGBM model used for interpretability analysis.

Table 1. Test performance of the selected models

Model	RMSE (Kw)	MAE (kW)	R ²
Weighted Ensemble	505.90	250.63	0.8579
LightGBM	509.32	250.32	0.8560

The best model selected by AutoGluon was Weighted Ensemble. This model combined the predictions of LightGBMXT, LightGBM, XGBoost, and Extra Trees base learners and achieved the best overall predictive performance on the chronological test set. This model achieved an RMSE of 505.90 kW, an MAE of 250.63 kW, and an R² value of 0.8579 on the chronological test set. This result indicates that the proposed model explains approximately 85.8% of the variation in measured active power using only instantaneous wind speed and wind direction information.

The LightGBM model achieved a very similar performance, with an RMSE of 509.32 kW, an MAE of 250.32 kW, and an R² value of 0.8560. Since its predictive performance was close to the ensemble model and it provides a tree-based structure suitable for interpretability analysis, LightGBM was selected for SHAP-based model explanation. The small performance difference suggests that LightGBM can serve as a reasonable interpretable proxy for analyzing the main feature-response patterns observed in the AutoML framework.

The obtained results are important because no lagged power values, rolling statistics, or theoretical power curve values were used as model inputs. Therefore, the models were not allowed to rely on previous power output or directly embedded theoretical power information. Instead, they learned the relationship between instantaneous wind conditions and active power output from the measured data.

5.2 Graphical Results

Figure 1 compares the actual and predicted active power values obtained using the Weighted Ensemble model on the chronological test set. The predicted series generally follows the temporal behavior of the measured power output. In particular, the model captures both low-production and high-production intervals. However, some deviations are observed during abrupt transitions and near-zero production periods. These deviations are expected because the model

uses only instantaneous wind speed and wind direction, without operational status variables, lagged power values, or turbine control information.

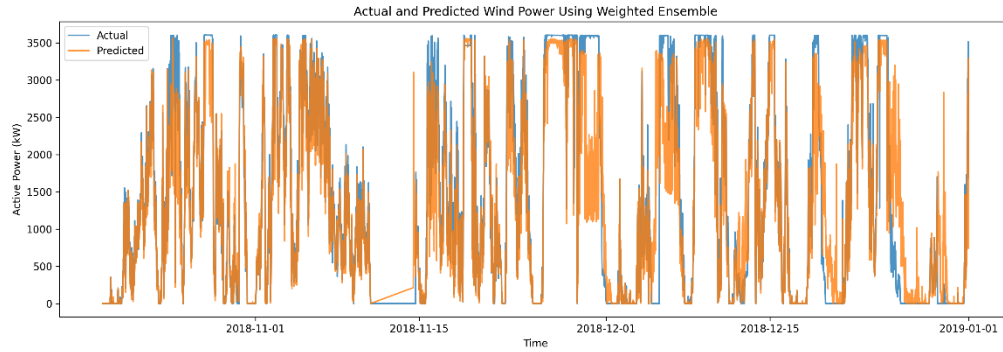


Figure 1. Actual and predicted wind power values using Weighted Ensemble

Figure 2 presents the actual and predicted active power values obtained using the LightGBM model. The graphical pattern is very similar to the ensemble model, which is consistent with the close numerical performance of the two models. This result supports the use of LightGBM for interpretability analysis, since it provides nearly comparable predictive accuracy while allowing a clearer examination of feature contribution patterns.

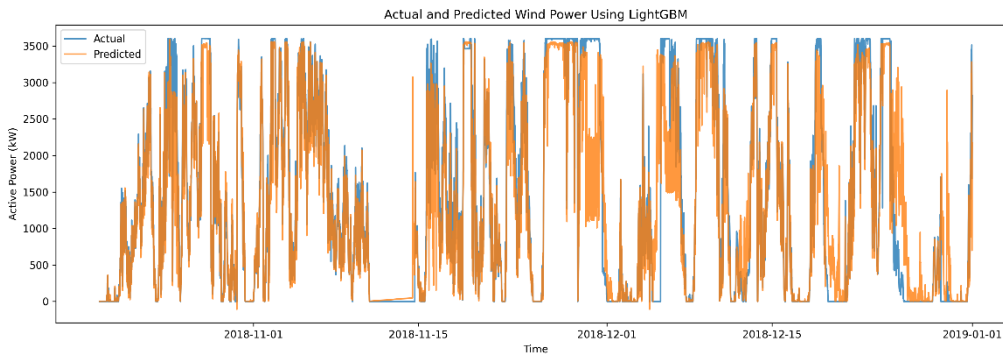


Figure 2. Actual and predicted wind power values using LightGBM

Figure 3 shows the SHAP summary plot for the LightGBM model. The results indicate that wind speed is the dominant input variable affecting the model output. Compared with wind speed, the sine and cosine components of wind direction have considerably smaller effects. This finding is physically meaningful, since wind power generation is primarily governed by wind speed, while wind direction has a secondary effect under the selected modeling setting.

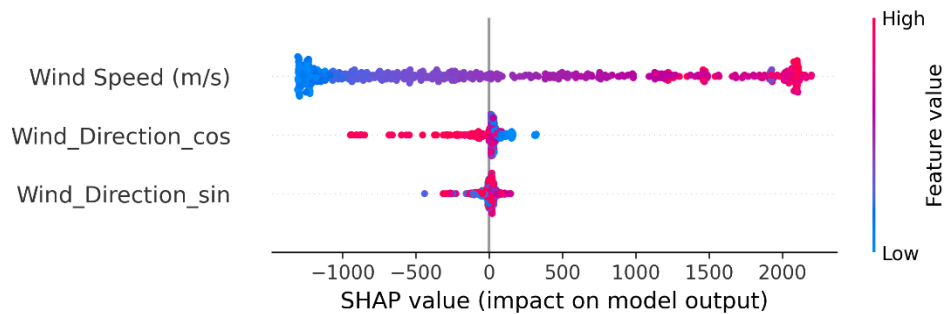


Figure 3. SHAP summary plot for the LightGBM model

Figure 4 illustrates the SHAP dependence of wind speed. The plot shows a nonlinear relationship between wind speed and its SHAP contribution. At low wind speeds, the SHAP values are negative, indicating a reducing effect on

predicted power output. As wind speed increases, the SHAP contribution rises sharply and becomes strongly positive. At higher wind speed levels, the contribution tends to plateau, which is consistent with the typical behavior of wind turbine power curves, where power output increases rapidly before approaching the rated power region.

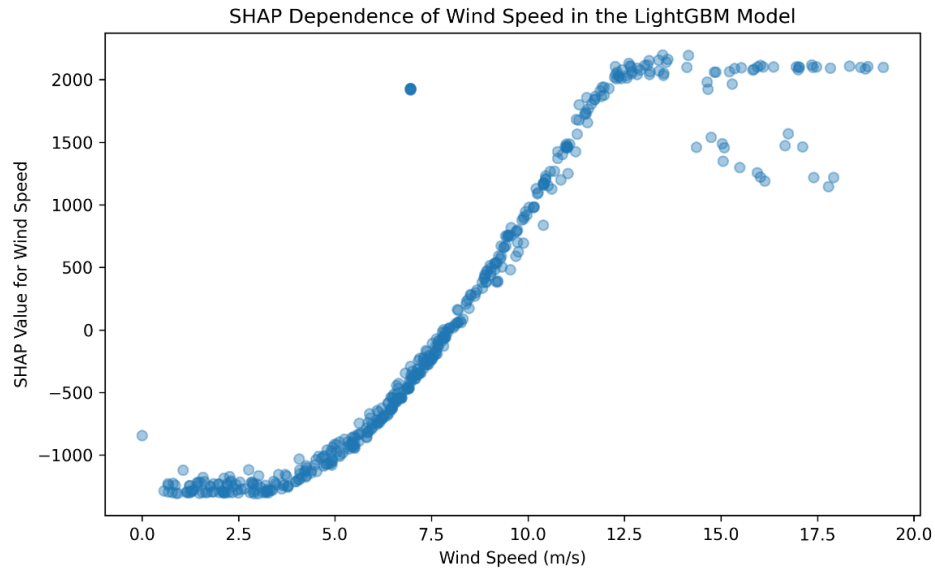


Figure 4. SHAP dependence plot of wind speed for the LightGBM model

Figure 5 presents the cubic polynomial fit applied to the wind speed SHAP contributions. The cubic fit achieved an R^2 value of 0.9231, indicating a strong agreement between the nonlinear SHAP pattern and a third-degree polynomial trend. This result supports the interpretation that the model learned a nonlinear wind speed–power relationship that is consistent with the expected physical behavior of wind power generation. However, this analysis should not be interpreted as a mathematical proof of causality; rather, it provides model-based evidence that the learned relationship is physically plausible.

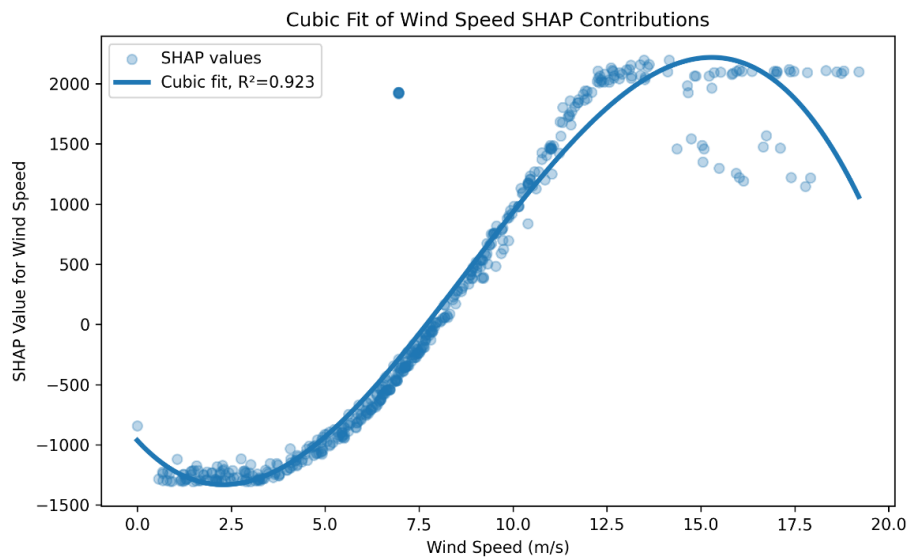


Figure 5. Cubic fit of wind speed SHAP contributions in the LightGBM model

5.3 Proposed Improvements

Although the proposed framework achieved a high explanatory performance using only instantaneous wind speed and wind direction, several improvements can be considered in future studies. First, operational variables such as turbine status, pitch angle, rotor speed, generator temperature, or maintenance indicators could be incorporated if available. These variables may help explain near-zero production periods and sudden deviations that cannot be fully captured by wind speed and wind direction alone.

Second, future models may include carefully designed lagged variables and rolling statistics. In the present study, lagged power values were intentionally excluded to evaluate the predictive capacity of instantaneous physical inputs. However, in operational forecasting applications, lagged wind speed, lagged power, and rolling averages may improve short-term prediction accuracy. If such variables are used, they should be constructed strictly from past observations to prevent temporal data leakage.

Third, additional validation strategies can be used to improve the robustness of the results. For example, rolling-origin evaluation or walk-forward validation can be applied to test whether model performance remains stable across different temporal segments. This would provide stronger evidence of generalization under changing seasonal and operational conditions.

Finally, the SHAP-based physical consistency analysis can be extended by comparing the learned SHAP pattern with the theoretical turbine power curve. This comparison may help distinguish whether the model follows the expected physical trend across the cut-in, rated power, and high wind speed regions.

5.4 Validation

The validation of the proposed framework was performed using a chronological hold-out test strategy. Unlike random splitting, chronological splitting preserves the temporal order of the observations and prevents future information from being used during model training. This is especially important for wind power prediction because random splitting may overestimate performance by mixing similar temporal patterns across training and test sets.

The final 20% of the dataset was kept completely unseen during model training and was used only for testing. The Weighted Ensemble model achieved an R^2 value of 0.8579 on this unseen test period, while the LightGBM model achieved an R^2 value of 0.8560. The close performance of these two models indicates that the obtained results are not limited to a single model structure.

Interpretability-based validation was also performed through SHAP analysis. The SHAP summary plot confirmed that wind speed was the most influential input variable. In addition, the SHAP dependence plot showed a nonlinear increasing pattern between wind speed and its contribution to predicted power. The cubic polynomial fit of the wind speed SHAP values achieved an R^2 value of 0.9231. This result supports the physical consistency of the learned model behavior, since wind power generation is theoretically expected to exhibit a nonlinear relationship with wind speed. Overall, the validation results show that the proposed framework provides both acceptable predictive performance and physically interpretable model behavior under a leakage-aware chronological test setting.

6. Conclusion

This study developed a leakage-aware AutoML-based framework for wind power prediction using only instantaneous wind speed and wind direction information. The dataset was split chronologically into training and test sets to prevent temporal data leakage. Outliers were treated using wind-speed-bin-specific IQR thresholds and median active power values derived only from the training set.

The best AutoGluon model, Weighted Ensemble, achieved an RMSE of 505.90 kW, an MAE of 250.63 kW, and an R^2 value of 0.8579 on the chronological test set. These results show that the proposed framework can explain approximately 85.8% of the variation in measured active power without using lagged power values or theoretical power curve information.

In addition to prediction accuracy, the study evaluated model interpretability through SHAP analysis. The LightGBM model was selected for this analysis because it achieved performance close to the ensemble model and provides a suitable tree-based structure for feature contribution analysis. The SHAP results showed that wind speed was the

dominant variable affecting power prediction. Furthermore, the cubic fit of wind speed SHAP contributions achieved an R^2 value of 0.9231, supporting the physical consistency of the learned nonlinear relationship.

The main contribution of this study is the combination of leakage-aware chronological evaluation, AutoML-based model comparison, and SHAP-based physical consistency assessment for wind power prediction. Future studies may improve the framework by incorporating operational turbine variables, walk-forward validation, and carefully constructed lagged features for short-term forecasting applications.

References

- Ahmad, T. and Zhang, D., A data-driven deep sequence-to-sequence long-short memory method along with a gated recurrent neural network for wind power forecasting, *Energy*, vol. 239, pp. 122109, 2022.
- Alkesaiberi, A., Harrou, F. and Sun, Y., Efficient wind power prediction using machine learning methods: A comparative study, *Energies*, vol. 15, no. 7, pp. 2327, 2022.
- Bergmeir, C. and Benítez, J. M., On the use of cross-validation for time series predictor evaluation, *Information Sciences*, vol. 191, pp. 192-213, 2012.
- Cakiroglu, C., Demir, S., Ozdemir, M. H., Aylak, B. L., Sariisik, G. and Abualigah, L., Data-driven interpretable ensemble learning methods for the prediction of wind turbine power incorporating SHAP analysis, *Expert Systems with Applications*, vol. 237, pp. 121464, 2024.
- Cerqueira, V., Torgo, L. and Mozetič, I., Evaluating time series forecasting models: An empirical study on performance estimation methods, *Machine Learning*, vol. 109, no. 11, pp. 1997-2028, 2020.
- Demolli, H., Dokuz, A. S., Ecemis, A. and Gokcek, M., Wind power forecasting based on daily wind speed data using machine learning algorithms, *Energy Conversion and Management*, vol. 198, pp. 111823, 2019.
- Erisen, B., Wind Turbine SCADA Dataset, Kaggle, Available: <https://www.kaggle.com/datasets/berkerisen/wind-turbine-scada-dataset>, Accessed on June 15, 2026.
- Han, S., Qiao, Y., Yan, P., Yan, J., Liu, Y. and Li, L., Wind turbine power curve modeling based on interval extreme probability density for the integration of renewable energies and electric vehicles, *Renewable Energy*, vol. 157, pp. 190-203, 2020.
- Liao, W., Fang, J., Ye, L., Bak-Jensen, B., Yang, Z. and Porte-Agel, F., Can we trust explainable artificial intelligence in wind power forecasting?, *Applied Energy*, vol. 376, pp. 124273, 2024.
- Loza, B., Minchala, L. I., Ochoa-Correa, D. and Martínez, S., Grid-friendly integration of wind energy: a review of power forecasting and frequency control techniques, *Sustainability*, vol. 16, no. 21, pp. 9535, 2024.
- Lundberg, S. M. and Lee, S. I., A unified approach to interpreting model predictions, *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- Muhammed Anees, V. V., Nazar, K. P. A., Sridharan, R. and Maniyath, S., Explainable and secure wind power forecasting: SHAP-enhanced framework with feature stability, anomaly detection, and adversarial robustness, *Applied Soft Computing*, vol. 196, pp. 115091, 2026.

Biographies

Tuğçe İnağ is a Research Assistant Dr. in the Department of Industrial Engineering at Gazi University, Ankara, Türkiye. She received her Ph.D. in Industrial Engineering from Gazi University. Her research interests include optimization, intelligent transportation systems, machine learning, forecasting, explainable artificial intelligence, and decision support systems. She has conducted studies on adaptive traffic signal control and artificial intelligence applications in engineering systems.

Büşra Bişkin is an Assistant Professor Dr. in the Department of Industrial and Systems Engineering at the Turkish Military Academy, National Defense University, Türkiye. Her research interests include vehicle routing, heuristics, multi-objective optimization, route planning, operations research, stochastic processes, and decision-making methods. She has conducted studies on routing and optimization problems, including heuristic approaches for multi-objective unmanned aerial vehicle routing.

Gül Didem Batur Sir is a Professor and the Department Vice Chair of Industrial Engineering at Gazi University, Türkiye. She received her B.S. and Ph.D. degrees from Gazi University and her M.S. degree from Bilkent University. Her research focuses on production planning and scheduling, optimization in manufacturing systems, and healthcare management. She has published in leading journals such as *International Journal of Production Economics*, *Computers*

*Proceedings of the 16th International Conference on Industrial Engineering and Operations Management
Ankara, Türkiye, June 23-25, 2026*

& Industrial Engineering, and Socio-Economic Planning Science, and has contributed to international engineering societies and conference organizations.