Robust Power Prediction of Wind Turbine using Error Detection, Clustering-Based Imputation and Physics-Informed Learning

Swayam Mittal, Vishwaas Narasinh, Nikhil Kulkarni,
Remish Leonard Minz, Nilanjan Chakravortty, Prateek Mital
Research & Development Hitachi India Ltd. Bangalore, India
Email:{swayam.mittal, vishwaas.narasinh, nikhil.kulkarni, remish.minz,
nilanjan.chakravortty, prateek.mital}
@hitachi.co.in

***Abstract*—In this paper, we present a robust power prediction model for wind turbines. Our model leverages error detection in the sensor data, clustering-based imputation of filtered erroneous or missing data, and a Physics-Informed Neural Network (PINN). We introduce data preprocessing steps, including the detection and filtering of erroneous data and clustering-based data imputation. We demonstrate that these preprocessing steps, along with the PINN framework, improve power prediction accuracy in the presence of erroneous sensor data.**

Keywords-wind farm, anomaly detection, power prediction, machine learning, clustering, physics-informed learning.

I. Introduction

Wind energy has carved a significant niche in today's renewable energy spectrum, offering a sustainable solution to the burgeoning global energy demands. With the increasing deployment of wind turbines, the volume of operational data they generate has surged, highlighting the necessity for advanced analytical techniques [1]. Safeguarding the integrity and precision of this data becomes imperative, particularly in the face of missing or erroneous readings [2]. While traditional solutions have earned recognition, they sometimes fall short of encapsulating the intricate dynamics of wind turbines [4]. Modern advancements lean towards sophisticated models, like auto-encoders, boasting improved accuracy [5]. However, an evident gap persists in ensuring these models align both with data-driven insights and inherent physical principles.

Power prediction of wind turbines faces challenges due to errors in the collected data. It is important to identify the erroneous data using anomaly detection methods to ensure power prediction accuracy. Further, once the detected erroneous data are filtered out, imputation of the filtered data and inherently missing data is required. Imputation may require advanced techniques to address the non-linear nature of the association between the longitudinal data. A particularly promising direction in addressing data imputation is the application of Gaussian Mixture Models (GMM). GMMs have demonstrated advantages in capturing complex data distributions, making them apt for handling the diverse nature of wind turbine data. This paper presents a novel pipeline, which includes error detection and filtering using the anomaly detection methods, clustering-based data imputation and physics-informed learning where we combine data-driven methods with physics-based predictions to address existing gaps.

The motivation for this work stems from the critical need to enhance the reliability and accuracy of power predictions in wind farms. As wind energy becomes a more significant component of the global energy mix, the ability to predict power output accurately under various operational conditions is essential for grid stability and efficient energy management. Traditional methods often fail to adequately address the complexities introduced by erroneous and missing data in wind turbine operations. Our approach aims to bridge this gap by integrating advanced data processing techniques with physics-informed models, thus providing a more robust and accurate power prediction framework.

The remainder of this paper is organized as follows: In Section II, we review related work in the fields of wind turbine data analysis and predictive modeling, emphasizing the significance of integrating data-driven approaches with physical models. In Section III, we detail the data collection process and the characteristics of the dataset used in this study. Section IV discusses our methodology for handling outliers and abnormalities in Supervisory Control and Data Acquisition (SCADA) data. Section V presents the methodologies used for anomaly detection and data imputation, followed by Section VI, where we present the description of our power output modeling approach. In Section VII, we evaluate the performance of our proposed models against various benchmarks and imputation techniques. Section VIII discusses the results of our experiments. Finally, Section IX concludes the paper with a summary of our findings and suggestions for future research.

II. Related Work

Research in wind turbine data analysis and prediction has been a burgeoning field over the past few years, with numerous methodologies developed to navigate the complexities posed by the vast datasets generated by wind turbines.

Errors in the sensor data pose major challenges in wind turbine power prediction. Various anomaly detection techniques are used for the detection of these erroneous data. [8] proposed an anomaly detection method based on a convolutional

recurrent autoencoder, showcasing the potential for leveraging deep learning models in this domain.

When the detected erroneous data are filtered from the dataset, they leave gaps in the dataset, making it all the more challenging to construct a correct power prediction model. Data imputation techniques are used to address this issue. [4] delved into traditional data imputation methods. Despite their widespread use, these methods have been found lacking imputation of data with non-linear associations, especially when applied to complex longitudinal data intrinsic to wind turbine operations. The oft-used strategy of substituting missing values with mean or median, as discussed by [2], can occasionally oversimplify the intricate interrelations inherent in turbine datasets.

More advanced techniques like auto-encoders for imputation in wind turbine sensor data have been explored by [5]. While their approach represents one of the latest advancements in data imputation techniques for wind turbines, however, there can be a large number of undetected erroneous data. In this situation, the correctness of the prediction model is questionable when we rely only on the data-driven approach. A promising avenue in addressing this shortcoming is the signals from the physics of the system in consideration. We leverage the use of Physics-Informed Neural Networks (PINNs). [9] presented a study on PINNs for power systems, emphasizing their capacity to integrate physical laws. Further building on this concept, [10] applied PINNs for non-linear system identification in power system dynamics, underlining the potential of combining data-driven models with physical insights.

Our study builds upon these foundational research endeavors. We aim to amalgamate data-driven insights with physics-informed models [7], ensuring that predictions are not only precise but also grounded in real-world operational frameworks.

Recent studies have also explored the use of hybrid models combining machine learning with physical modeling. For instance, [11] proposed a hybrid model integrating a deep neural network with a physical wind model, showing improved accuracy in wind power prediction. Similarly, [12] introduced an ensemble learning approach that combines multiple machine learning models to enhance prediction robustness.

In comparison, our approach integrates Gaussian Mixture Models for imputation and Physics-Informed Neural Networks to ensure that the predictions are not only accurate but also physically plausible. Unlike [11] and [12], which focus primarily on the data-driven aspects, our method emphasizes the integration of physical principles to handle erroneous and missing data more effectively.

Despite the advances in these techniques, several limitations persist in the state-of-the-art methods. Traditional anomaly detection and imputation methods often fail to account for the complex, non-linear relationships in wind turbine data, leading to suboptimal power prediction accuracy. Advanced methods such as auto-encoders improve upon these issues but still suffer from undetected anomalies and reliance on purely data-driven approaches, which may not fully capture the physical dynamics of wind turbines. The integration of physics-informed models, while promising, also presents challenges in terms of model complexity and computational requirements. Our work seeks to address these limitations by providing a comprehensive framework that combines robust data preprocessing, advanced imputation techniques, and physics-informed neural networks to enhance power prediction accuracy and reliability.

III. Data Collection Overview

The wind farm under consideration is an onshore wind farm, built in 2017–2018 and has been operating since 2019. For this investigation, the turbine data was collected between November 1st, 2022 and July 15th, 2023. There are 16 wind turbines with a total of 32MW power generation capacity. The wind farm had access to a collection of 1966 SCADA tags that contained information from various turbine components, including the rotor, brake, pitch control, main shaft, gearbox, generator, yaw system, nacelle, electrical systems, hydraulic systems, etc. Our focus is on two wind turbines. The data samples from the sensors (SCADA parameters) are averaged across a 10-minute timeframe and are recorded at a frequency of roughly 2.00 min. We filtered the data for the wind turbine operational phase, focused on the core aspects of the wind turbine which are power generation, rotor, and pitch, and removed outlier records with very low wind speed (< 3 m/s) and very high wind speed (> 10.5 m/s), or records with no production (0 kWh).

The wind speed classification mentioned above is based on the manufacturer's specifications. However, it is essential to note that specific wind turbine models might have slightly different operational parameters.

IV. Handling Outliers and Abnormality in SCADA Data

A. Outliers Observed

Outliers in SCADA data can greatly influence the performance and accuracy of wind turbine predictive models. In this study, several types of outliers were observed:

- **Non-operating Phase Outliers:** During wind turbine maintenance, the value of the active power is zero even though the wind speed lies between the cut-in and cut-out speeds. These data points were identified as part of the non-operating phase and were systematically removed to avoid misinterpretation.
- **Power Curve Deviation Outliers:** Some data points, although not zero, were observed to deviate significantly from their expected values on the power curve. Probable causes for such deviations include wind curtailment, accumulation of dirt or bugs on the turbine blades, pitch malfunctions, among other operational issues.

B. Methodology for Handling Outliers

The approach adopted to address the identified outliers involved the following steps:

1) **Initial Identification:** A visual examination was first conducted on plots of wind speed versus wind power. This helped in identifying data points that significantly deviated from the expected behavior.
2) **Interval-based Detection:** Following the initial identification, we employed the interval-based detection method as described by [8]. This method allows for the removal of obvious outliers based on set intervals or thresholds in the plot of wind speed and wind power. Specifically, data points that fall outside of expected performance intervals were flagged.
3) **Power Curve Validation:** Given the inherent relationship between wind speed and turbine output power, we used the power curve as a benchmark. Any data points that strayed significantly from the power curve were considered outliers. This step was particularly useful for identifying the Power Curve Deviation Outliers.

However, given the specific nature of wind turbine data, we decided to rely more on domain knowledge for this study. Once outliers were identified through the above methodologies, they were systematically removed from the dataset. Following the removal of these outliers, the refined wind data were employed to train the power curve models. It is worth noting that a meticulous outlier removal process ensures the developed models' robustness and accuracy in predicting wind turbine performance based on SCADA data.

V. Methodology

The overall methodology of our study is depicted in Figure 1. This flow diagram outlines the primary steps involved in the data processing and analysis phases.

The flow diagram above provides a visual summary of our approach, including the key steps and processes involved.

A. Outlier Handling

For handling outliers detected during non-operational or maintenance phases, we visually inspected for wind speed versus power plots. Leveraging the domain expertise, we identified and eliminated these outliers, enhancing the data quality for subsequent model training. Detailed data collection processes are discussed in the Data Collection Overview.

B. Feature Selection

Our study utilizes SCADA data obtained from a real-world wind farm situated in Gujarat, India. Spanning a specific timeframe, this dataset offers insights into various turbine components, painting a comprehensive picture of the turbine operations.

Initial feature selection was guided by a combination of domain knowledge, data availability, and feature importance scores. Starting with a broad set of SCADA tags, domain expertise helped shortlist a preliminary set of 80 features. Additionally, essential parameters like wind speed, rotor speed, and pitch angles were mandated by the physics loss function.

To further refine our features list, a Random Forest model was trained using 3-fold cross-validation, and the results

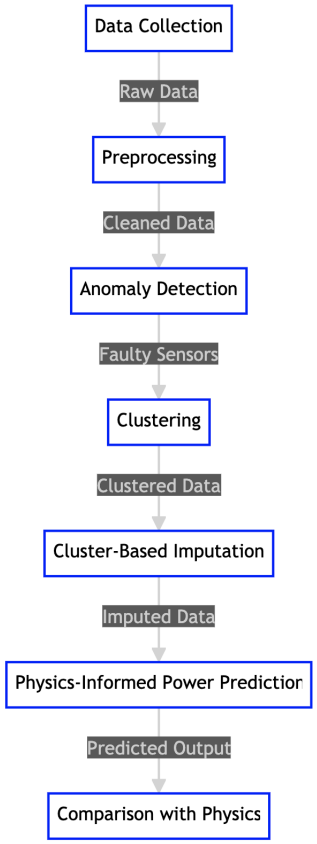

Figure 1. Flow diagram.

yielded the feature importance as demonstrated in Figure 2. From these, the top 10 features were selected for anomaly detection. The figure illustrates the top 5 features with the highest significance. The remaining features, while essential, have lesser importance values and are not prominently displayed in the graph.

The significance of each feature used in our models is illustrated in Figure 2.

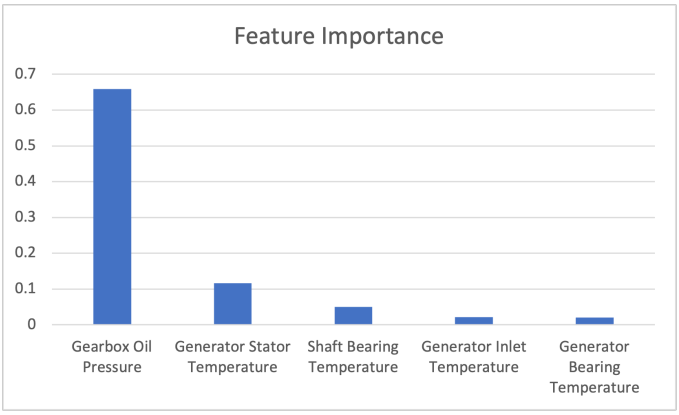

Figure 2. Feature Importances obtained from Random Forest model.

The final set of top 10 features includes:

- Gearbox Oil Pressure
- Generator Stator Temperature
- Shaft Bearing Temperature
- Generator Inlet Temperature
- Generator Bearing Temperature
- Pitch Angle
- Wind Speed
- Rotor Speed
- Nacelle Direction
- Yaw

C. Anomaly Detection

We attempt to identify faulty sensors for anomaly detection using auto-encoders. Consequently, the inputs and the outputs of the auto-encoder are the same. The architecture of the auto-encoder is shown in Figure 3. The input features include the features discussed above.

An auto-encoder is a type of artificial neural network that can learn efficient representations of input data with no need for labels. It consists of two parts: an encoder that compresses the input into a latent-space representation, and a decoder that reconstructs the input from this representation. The goal is to minimize the difference between the input and the reconstructed output.

In Figure 3, the auto-encoder architecture is detailed as follows:

- **Input Layer (input_1):** This layer accepts the input data with shape (None, 3, 11), where 'None' represents the batch size, 3 represents the sequence length, and 11 represents the number of features.
- **Conv1D Layer (conv1d):** This layer applies 1D convolution to the input data, reducing the feature dimension from 11 to 4.
- **Dropout Layer (dropout):** This layer randomly sets a fraction of input units to 0 to prevent overfitting.
- **Conv1D Layer (conv1d_1):** Another convolutional layer that further reduces the feature dimension to 1.
- **Conv1DTranspose Layer (conv1d_transpose):** This transposed convolutional layer starts the decoding process, increasing the feature dimension back to 4.
- **Dropout Layer (dropout_1):** Another dropout layer to prevent overfitting during the decoding process.
- **Conv1DTranspose Layer (conv1d_transpose_1):** The final transposed convolutional layer reconstructs the output to match the original input shape of (None, 3, 11).

We train the model in 3-fold cross-validation for 10 epochs using Mean Absolute Error (MAE) loss between the target and the prediction. The distribution of the training loss is shown in Figure 5. We set the threshold at the 90th percentile which corresponds to a value of 0.15. This decision is based on empirical observations to capture the most significant anomalies while reducing the likelihood of false positives. Consequently, any loss greater than the defined threshold is considered to be an anomaly.

For testing, we randomly select a couple of features, change their value to $\mu_i \pm 2\sigma_i$, and keep the other features at their mean value. i represents the selected features. The output of the anomaly detection model with this anomalous input is subtracted from the mean of the output of the training data to get the loss due to the anomaly. This is shown in Figure 6, where the peaks corresponding to the anomalous features have a higher loss and have crossed the threshold defined above.

However, we only focus on the cases where either one of the sensors that correspond to wind speed, rotor speed, or pitch angle is at fault. This is because the empirical relation to the expected power output and the physics loss functions require these features to be present.

The notation $\mu \pm 2\sigma$ is conventionally used to describe data lying within two standard deviations (σ) from the mean (μ). In a Gaussian distribution, roughly 95.4% of data falls within this range. Before applying this principle to identify outliers, we verified that our features adhere to a Gaussian distribution, with various statistical methods. This validation ensures the appropriateness of the $\mu \pm 2\sigma$ rule in our context.

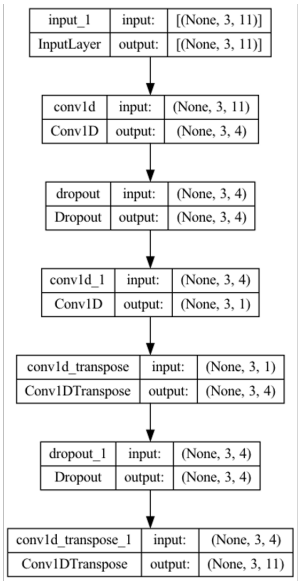

Figure 3. Auto Encoder architecture.

D. Outlier Detection using Standard Deviation

In the process of data pre-processing, it is crucial to identify and handle outliers that can influence the outcomes of the analysis. One effective method employed in this study involves the use of standard deviation.

Given a dataset, the mean (μ) represents the average value, while the standard deviation (σ) provides a measure of the data's spread or dispersion. In a normally distributed dataset, approximately 68.2% of the data lies within $\mu \pm \sigma$, and about 95.4% lies within $\mu \pm 2\sigma$. Data points that fall outside of $\mu \pm 2\sigma$ can be considered as potential outliers, as they deviate significantly from the mean.

In our analysis, data points falling outside the range of $\mu \pm 2\sigma$ were further investigated to determine their validity and were treated or removed accordingly.

E. Clustering

GMMs are chosen to cluster turbines based on multiple features due to their capacity to model complex data distributions. The features selected for clustering are the ones previously mentioned, except for wind speed, rotor speed, and pitch angles. To determine the optimal number of clusters for the GMM, we employ the elbow method, visually represented in Figure 4.

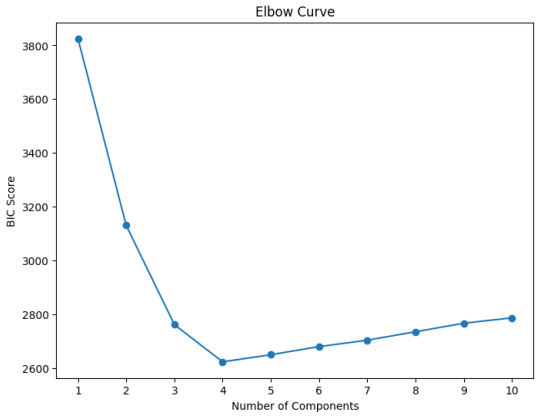

Figure 4. Optimal number of k clusters using the elbow method.

Recognizing the importance of data quality, we introduce a clustering-based imputation methodology. GMMs, with their probabilistic framework, offer an advantage over deterministic clustering methods like K-means. Using the GMM, we cluster wind turbines based on operational and spatial parameters, allowing for effective imputation of missing values. The clustering phase involves grouping wind turbines, guided by important features derived from the Random Forest model. By training the GMM on these scaled features, we ensure uniform scaling and compatibility.

During the validation phase, the test dataset is meticulously constructed with Gaussian distribution to encompass wind turbine feature values that replicate diverse operational scenarios. Here, faulty sensor readings are replaced with the mean values of their corresponding clusters. The findings are compelling, as we observed a close match between the imputed values and the actual expected values across various test scenarios, substantiating the imputation mechanism's accuracy.

Although GMMs come with their assumptions, especially about cluster shapes, and can be sensitive to initialization, we chose to use GMMs because of their strengths and the specific characteristics of our dataset.

The distribution of training loss across various thresholds, which is critical for setting our anomaly detection parameters, is presented in Figure 5.

Figure 6 shows the results of our anomaly detection process, highlighting how our model responds to different types of sensor errors.

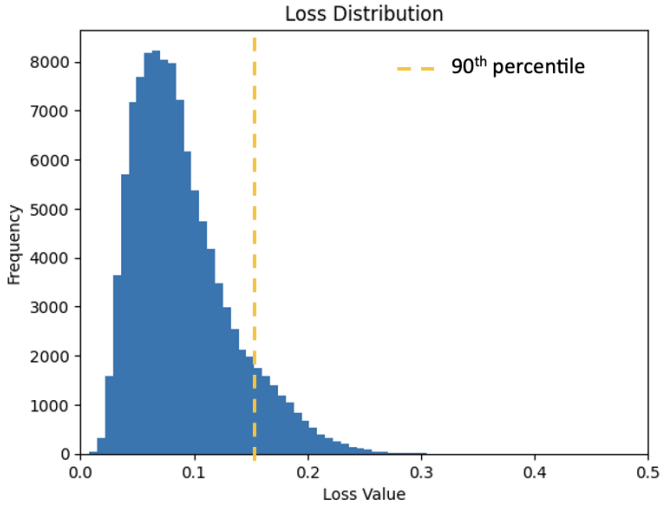

Figure 5. Loss distribution and threshold showing how losses are distributed across different thresholds.

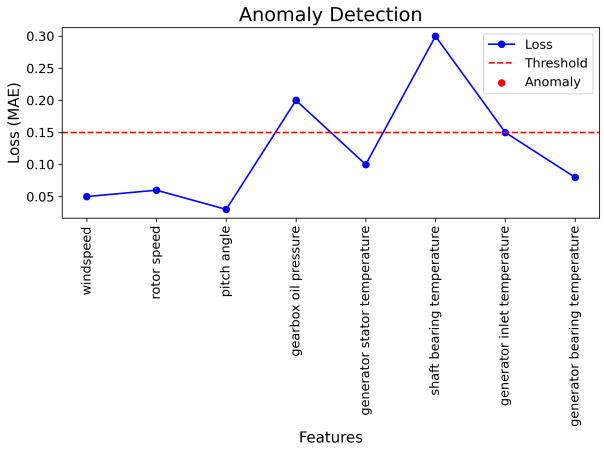

Figure 6. Anomaly detection results.

VI. Power Output Modelling

Our power prediction model, built atop this preprocessed data, comprises multiple layers of fully connected neural networks. The model's architecture, training configurations, and hyperparameters are detailed in this section.

With the imputed values for the faulty sensor data, we model the power output of the wind turbine using the features mentioned above. The model consists of 4 layers of fully connected neural networks with Rectified Linear Unit (ReLU) activation units and dropouts between each layer.

During the training phase, we maintain the actual values for all features, i.e., it is trained with non-faulty sensors in 3-fold cross-validation for 30 epochs, which is approximately the number of steps during which the validation loss stabilizes. During the validation and testing phase, features other than wind speed and rotor speed are fed to the clustering algorithm, which clusters the instances into a cluster. The faulty sensor value is replaced by taking the mean of the actual sensor values

from the training data for the selected cluster. We select the best model based on the validation loss.

A. Physics-Informed Loss

The crux of our approach lies in integrating a physics-based loss function. We derive this loss from the energy conservation laws governing wind turbines, ensuring our model's predictions are both data-driven and physically informed. Although the model with traditional MAE loss function converges, there is often a need to include physical laws in the system. We incorporate physics into our model via loss functions. The physical laws are derived using energy conservation laws at different stages of the turbine, as shown in Figure 7.

The stages of power loss throughout the wind turbine system are highlighted in Figure 7. This diagram assists in understanding the energy flow and losses at various stages.

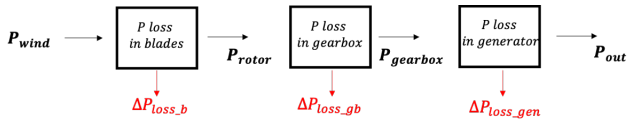

Figure 7. Power Loss at various stages.

The power in the wind is given by [6], where A is the area swept by the blades, v is the wind velocity, ρ is the air density.

$$P_{wind} = \frac{1}{2}\rho A v^3 \tag{1}$$

There are various methods to model the power coefficient, which is a function of the tip speed ratio of the rotor blade, β and the pitch angle, λ, thus $C_p(\lambda, \beta)$. Reyes et at. present a review in [6], stating there are three major approaches to model C_p, namely, the polynomial model, the sinusoidal model and the exponential model. The names indicate how the general function is used to model C_p from either the tip speed ratio λ or both the tip speed ratio λ and the pitch angle β. We use a generic formulation of the widely adopted exponential model from [6] rewritten in equation (2) and (3). The typical values of the coefficients used in both of these equations are described in table 7 and table 8 in [6]. In our implementation, we use the most widely used exponential model's coefficient values, as described in [13].

$$C_p = c_0(c_1\lambda_i^{-1} + c_2\beta + c_3\beta^{c_4} + c_5)e^{c_6\lambda_i^{-1}} + c_7\lambda \tag{2}$$

$$\frac{1}{\lambda_i} = \frac{1}{\lambda + d_0\beta + d_1} - \frac{d_2}{1+\beta^3} \tag{3}$$

The power loss at the blades is given by equation (4). The power loss at the gearbox is given by equation (5). η_{gb} represents the efficiency coefficient of the turbine gearbox. Similarly, the loss at the turbine generator is given by equation (6), where η_{gen} represents the turbine generator's efficiency coefficient. These efficiency values are evaluated using an iterative method, as described in [14].

$$\Delta P_{loss\_b} = (1 - C_p)P_{wind} \tag{4}$$

$$\Delta P_{loss\_gb} = (1 - \eta_{gb})P_{rotor} \tag{5}$$

$$\Delta P_{loss\_gen} = (1 - \eta_{gen})P_{gearbox} \tag{6}$$

The power output accounted for the above losses is given in equation (7). Simplifying equation (7) expressed the power output in terms of the power coefficient, the gearbox efficiency coefficient and the generator efficiency coefficient, presented in equation (8).

$$P_{out} = P_{wind} - (\Delta P_{loss\_b} + \Delta P_{loss\_gb} + \Delta P_{loss\_gen}) \tag{7}$$

$$P_{out} = C_p\eta_{gb}\eta_{gen}P_{wind} \tag{8}$$

The physics loss is given by the difference between the predicted output power of the wind turbine and the actual power produced by the wind turbine for a stipulated period. This is expressed in equation (9), where P_{out} denotes the predicted output power of the wind turbine and P_{actual} denotes the actual power produced by the turbine.

$$Loss_{physics} = P_{out} - P_{actual} \tag{9}$$

TABLE I
COMPARISON OF IMPUTATION ACCURACY USING MEAN ABSOLUTE ERROR (MAE) FOR GMM, K-MEANS, AUTOENCODER, AND SIMPLE AVERAGE METHODS

Method	MAE (Imputation Accuracy)
GMM	9.21
Autoencoder	12.62
KMeans	15.86
Simple Average	32.69

Table I summarizes the efficacy of different imputation methods using the Mean Absolute Error (MAE) metric. The GMM-based method shows the lowest MAE, indicating better imputation accuracy compared to the K-means, Autoencoder, and Simple Average methods.

Figure 8 illustrates the validation of introduced anomalies. The outcomes help verify the sensitivity of our anomaly detection system.

VII. EVALUATION

The evaluation of our proposed methodology focuses on two primary goals: 1) Assessing the accuracy of the power prediction model and 2) Validating the robustness of the model against faulty sensor data and imputation techniques.

A. Evaluation Methodology

To achieve these goals, we conducted a comprehensive set of experiments involving the following steps:

1) **Data Preprocessing:** This step includes outlier detection and handling, feature selection, and anomaly detection, as described in the Methodology section.

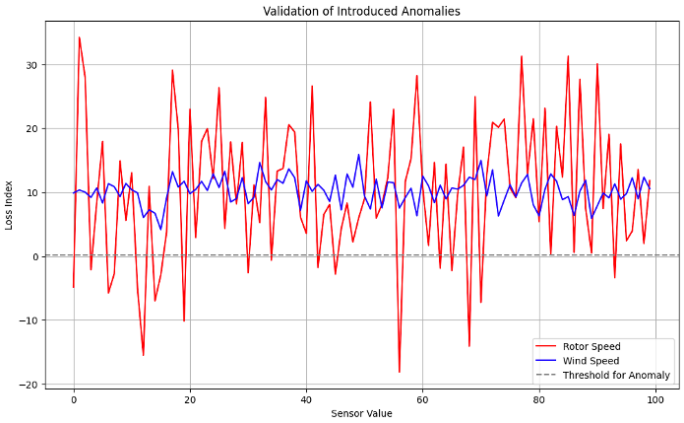

Figure 8. Validation of Introduced Anomalies.

2) **Imputation Techniques Comparison:** We evaluated various imputation techniques, including Gaussian Mixture Models (GMM), Autoencoder-based imputation, K-means clustering, and Simple Average imputation. The effectiveness of these techniques was assessed using the Mean Absolute Error (MAE) metric.
3) **Model Training and Validation:** The power prediction model was trained using the preprocessed and imputed data. We used a 3-fold cross-validation approach to ensure the robustness of the model. The model architecture included multiple layers of fully connected neural networks with ReLU activations and dropout regularization.
4) **Physics-Informed Neural Networks (PINN):** We integrated a physics-based loss function into the neural network to align the predictions with physical laws governing wind turbines. The impact of this integration was evaluated by comparing the performance of models with and without the physics loss.
5) **Benchmarking Against State-of-the-Art:** We benchmarked our model against traditional power prediction models and recent advancements such as autoencoder-based methods.
6) **Ablation Study:** To further understand the contribution of each feature, we conducted an ablation study where each feature was adjusted from its mean value and the prediction accuracy was observed with and without the physics model.

B. Main Goals

The main goals of our evaluation are as follows:

- **Accuracy of Power Prediction:** Determine the accuracy of our power prediction model by comparing predicted power outputs with actual values, using metrics such as Mean Absolute Error (MAE) and the coefficient of determination (R^2).
- **Robustness of Imputation Techniques:** Validate the effectiveness of the GMM-based imputation technique compared to other methods, especially in handling non-linear associations in the data.
- **Impact of Physics-Informed Learning:** Evaluate the contribution of physics-informed neural networks in improving the prediction accuracy and ensuring that the model's predictions adhere to physical principles.
- **Comparative Analysis:** Benchmark the proposed methodology against state-of-the-art approaches to highlight the improvements and advantages of our integrated approach.
- **Feature Contribution:** Through the ablation study, assess the significance of individual features on the model's performance and demonstrate the necessity of combining domain-specific features with data-driven techniques.

The results from these evaluations are discussed in the subsequent section.

VIII. Results

To ensure a comprehensive benchmark, we compared the results of GMM-based clustering with methods [4] that use K-means clustering and also evaluated the two-stage deep autoencoder-based method, as proposed by [5]. Their approach primarily utilizes a deep autoencoder to recover the underlying structure of the data and then imputes the missing value. The proposed method using GMM shows a lower Mean Absolute Error (MAE), as shown in Table I. Our GMM-based clustering not only demonstrates a significant improvement over traditional K-means clustering but also outperforms the recent deep autoencoder-based method in terms of MAE.

For Anomalous sensor value detection, we benchmark the performance by changing the sensor value to various variations, as shown in Figure 8. We see that the induced variations in sensor values result in a loss well above the threshold for most variations of values.

Figure 9 displays the validation loss of our power prediction models over epochs, comparing models with and without the incorporation of physics-informed loss. This graphical representation helps in understanding the impact of physics-based modeling on the convergence and performance of the predictive models.

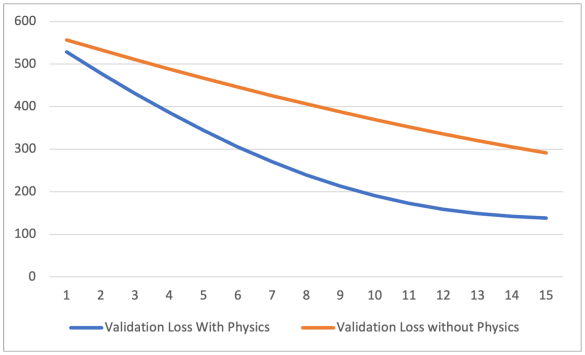

Figure 9. Validation Loss with and without physics loss. The X-axis represents the number of epochs, and the Y-axis represents the validation loss.

We benchmark the performance of our proposed methodology against traditional power prediction models. Both qualitative and quantitative analyses emphasize the advantages of our physics-informed approach. We obtain the results for the power prediction model with and without physics loss. We use the coefficient of determination R^2 as our metric to evaluate the performance of our model. The R^2 with physics loss seems to perform better, as shown in Table II.

We further investigate the convergence of the model with and without physics loss. We train the model in 5-fold cross-validation and aggregate the results and validation loss. We see that the model converges quicker with physics loss, as shown in Figure 9. Our imputation methodology significantly enhances the reliability of predictions. Furthermore, the integration of the physics-informed loss ensures our predictions are not just accurate but also adhere to the physical laws of wind turbine operations.

TABLE II
R^2 OF THE POWER PREDICTION MODEL WITH AND WITHOUT PHYSICS LOSS

Faulty sensor	Without Physics Loss	With Physics Loss
Wind Speed	0.67	0.77
Rotor Speed	0.51	0.58
None	0.77	0.79

TABLE III
ABLATION STUDY SHOWING THE ACCURACY (ACC) OF PREDICTIONS FOR DIFFERENT DEVIATIONS OF EACH FEATURE WITH AND WITHOUT THE PHYSICS (PHY) MODEL

Feature	Deviation	Acc w/o Phy	Acc w/ Phy
Gearbox Oil Pressure	+/- 1%	93.2%	94.1%
Gen. Starter Temp.	+/- 0.7%	92.5%	93.8%
Shaft Bearing Temp.	+/- 0.6%	93.0%	93.5%
Gen. Inlet Temp.	+/- 0.7%	92.8%	93.7%
Gen. Bearing Temp.	+/- 0.6%	93.1%	93.9%
Pitch Angle	+/- 0.5%	92.4%	93.3%
Wind Speed	+/- 2%	91.9%	93.0%
Rotor Speed	+/- 1%	93.2%	94.2%
Nacelle Direction	+/- 0.8%	92.6%	93.6%
Yaw	+/- 0.5%	92.3%	93.2%

To further evaluate our model's robustness to changes in input features, we performed an ablation study. In this study, we adjusted each feature from its mean value and examined the model's prediction accuracy, both with and without the use of the physics model. The findings are outlined in Table III. The table shows that using the physics model consistently improves prediction accuracy across all features. This underscores the importance of combining real-world physical knowledge with data-driven modeling. Features like 'Wind Speed' and 'Pitch Angle', which significantly influences turbine performance, benefits notably from the physics model. This supports the idea of using a physics-based modeling approach.

This study also indicates that our model can handle variations in data, making it suitable for real-world wind farm scenarios.

IX. CONCLUSION

This paper presents a comprehensive approach to wind turbine power prediction. Ensuring data quality through clustering-based imputation and integrating Physics-Informed Neural Networks for power prediction, we ensure that predictions are both accurate and physically feasible. Our methodology, tested on real-world data, underscores the importance of merging data-driven insights with domain-specific expertise, paving the way for future innovations in wind turbine operations and maintenance.

While our approach advances wind turbine power prediction, future work can optimize clustering for enhanced imputation, integrate real-time data with advanced neural architectures, and expand the method's applicability to other renewables like solar, ensuring data-driven yet physically coherent predictions.

REFERENCES

[1] J. Tautz-Weinert and S. Watson, "Using SCADA data for wind turbine condition monitoring—A review," IET Renew. Power Gener., vol. 10, no. 4, pp. 382–394, Sep. 2017.

[2] R. Razavi-Far and M. Saif, "Imputation of missing data for diagnosing sensor faults in a wind turbine," in Proc. IEEE Int. Conf. Syst., Man, Cybern., Hong Kong, Oct. 2015, pp. 99–104.

[3] B. Zhao, Y. Zhong, A. Ma, and L. Zhang, "A spatial Gaussian mixture model for optical remote sensing image clustering," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 9, no. 12, pp. 5748–5759, Dec. 2016.

[4] M. Morshedizadeh, M. Kordestani, R. Carriveau, D. S.-K. Ting, and M. Saif, "Application of imputation techniques and adaptive neuro-fuzzy inference system to predict wind turbine power production," Energy, vol. 138, pp. 394–404, Nov. 2014.

[5] X. Liu and Z. Zhang, "A two-stage deep autoencoder-based missing data imputation method for wind farm SCADA data," IEEE Sensors, vol. 21, no. 9, pp. 10933-10945, May 2021.

[6] V. Reyes, J. J. Rodriguez, O. Carranzo, and R. Ortega, "Review of mathematical models of both the power coefficient and the torque coefficient in wind turbines," 2015 IEEE 24th International Symposium on Industrial Electronics (ISIE), pp. 1-5.

[7] B. Huang and J. Wang, "Applications of Physics-Informed Neural Networks in Power Systems - A Review," IEEE Transactions on Power Systems, vol. 38, no. 1, pp. 572-588, Jan. 2023.

[8] C. Yin, S. Zhang, J. Wang, and N. N. Xiong, "Anomaly Detection Based on Convolutional Recurrent Autoencoder for IoT Time Series," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 52, no. 1, pp. 112-122, Jan. 2022.

[9] G. S. Misyris, A. Venzke, and S. Chatzivasileiadis, "Physics-Informed Neural Networks for Power Systems," 2020 IEEE Power & Energy Society General Meeting (PESGM), Montreal, QC, Canada, 2020, pp. 1-5.

[10] J. Stiasny, G. S. Misyris, and S. Chatzivasileiadis, "Physics-Informed Neural Networks for Non-linear System Identification for Power System Dynamics," 2021 IEEE Madrid PowerTech, Madrid, Spain, 2021, pp. 1-6.

[11] Z. Ma and G. Mei, "A hybrid attention-based deep learning approach for wind power prediction," Energy Reports, vol. 7, 2021, pp. 1461-1472. doi: https://doi.org/10.1016/j.egyai.2022.100199.

[12] J. Lee, W. Wang, F. Harrou, and Y. Sun, "Wind Power Prediction Using Ensemble Learning-Based Models," IEEE Transactions on Sustainable Energy, vol. 12, no. 2, 2021, pp. 1017-1027. doi: 10.1109/TSTE.2020.3042578.

[13] L. Lu, Z. Xie, X. Zhang, S. Yang and R. Cao, "A dynamic wind turbine simulator of the wind turbine generator system", International Conference on Intelligent System design and engineering application,pp. 967–970. DOI: 10.1109/ISdea.2012.549.

[14] J. Tamura, "Calculation Method of Losses and Efficiency of Wind Generators", Wind Energy Conversion System, pp. 25-51, January 2012.