e-Prime - Advances in Electrical Engineering, Electronics and Energy 13 (2025) 101043

Contents lists available at ScienceDirect

e-Prime - Advances in Electrical Engineering, Electronics and Energy

journal homepage: www.elsevier.com/locate/prime

Hybrid model for cleaning abnormal data of wind turbine power curve based on machine learning approaches

Abdelwahab Ayash Subuh [a,*], S. Hr. Aghay Kaboli [a], Muhammad Waqar [b], François Vallée [a]

[a] Power Systems and Markets Research Group, University of Mons, 7000 Mons, Belgium
[b] State company of electricity production/Northern region, Iraq

ARTICLE INFO

Keywords:
Wind turbine power curve
Outliers
Fuzzy c-means clustering
Mahalanobis distance
Artificial neural networks
Support vector regression
ANFIS
K-means
RMSE
MAPE

ABSTRACT

This paper addresses important challenges in wind energy prediction caused by outliers in wind data, which distort the wind turbine power curve and lead to inaccurate performance assessments and suboptimal operation strategies. The major difficulty here is detecting and eliminating these outliers from complex wind datasets, as inaccurate data can significantly impact forecasting and related activities. To overcome this challenge, the paper proposes a hybrid model combining fuzzy C-means clustering, Mahalanobis distance, and Artificial Neural Networks (ANN) to detect and remove outliers far more accurately than any individual method or other traditional hybrid method, decreasing false alarms and misses. It improves data quality and boosts the reliability of turbine performance analysis, resource assessment, and forecasting, supporting more efficient and sustainable wind-power operations. The results show (1) that the proposed hybrid model achieves 15.4 % more accuracy than the other traditional hybrid models in detecting and removing outliers. (2) The proposed hybrid model gives an overall ≈ 116.1 % improvement in outlier-detection accuracy over the individual models. (3) Adding the ANN to the proposed hybrid model boosts the outlier-detection accuracy to about a 69.5 % relative improvement. (4) Detecting and cleaning outliers by the proposed hybrid model cuts the RMSE from 2.38 to 1.27, reducing prediction error by 46.6 %. (5) The advanced hybrid model used in this study for comparison purposes achieves nearly identical accuracy to the proposed hybrid model; it reduces RMSE by ~0.015 and MAPE by ~0.04 pp and boosts R^2 by ~0.001 while maintaining almost perfect outlier detection (99 % vs. 100 %). Although the advanced model offers a marginal edge in reconstruction quality, the lightweight, scalable proposed hybrid model remains better appropriate for real-world deployment due to its lower computational overhead and more straightforward maintenance.

1. Introduction

1.1. Context

Wind data precision is essential in the energy industry for recreate wind turbine power curves. The power curve shows the connection between wind speed and turbines power generation play crucial role in evaluate performance, enhance operations and project energy production. Accurate wind measurements produce accurate power curves for supervising and caring for wind turbines. For instance, precise power curve models are instrumental in wind turbine selection, capacity factor determination, wind energy analysis and prediction, and condition monitoring [1]. Precise wind power predictions achieve a balance between power supply and demand, reduce reliance on backup power sources and lower overall energy production costs. Therefore, accurate wind data are vital for creating reliable wind turbine power curves and evaluating performance and energy output efficiently.

However, raw wind data may include outliers, data values that are considerably different from the rest. They may occur due to instrument errors, environmental interferences, or data transmission errors. If

Abbreviations: WTPC, wind turbine power curve; SCADA, supervisory control and data acquisition; FCM, fuzzy c-means clustering; ANFIS, adaptive neuro-fuzzy interference system; ANN, artificial neural network; SVR, support vector regression; MAE, mean absolute error; RMSE, root mean square error; R^2, coefficient of determination; MAPE, mean absolute percentage error; DAE, denoising autoencoder; LSTM A, long short-term memory autoencoder; DBSCAN, density-based spatial clustering of applications with noise.

* Corresponding author.
E-mail addresses: abdelwahab.SUBUH@student.umons.ac.be (A.A. Subuh), kaboli0004@gmail.com (S.Hr.A. Kaboli), francois.vallee@umons.ac.be (F. Vallée).

https://doi.org/10.1016/j.prime.2025.101043
Received 21 February 2025; Received in revised form 6 May 2025; Accepted 9 June 2025
Available online 10 June 2025

outliers are not correctly detected and handled, they can cause harm to applications of wind power predicting, power flow studies, and economic dispatch analysis. Lower accuracy, higher predicting errors, and suboptimal unit commitment findings are likely to be exhibited by models with outliers [2]. Hence, it is important to have a strong approach to identifying and dealing with outliers to ensure the reliability of the wind data analysis.

Traditional statistical analyses, including thresholding and z-score-based techniques, are popular but may not be very efficient in detecting anomalies in large datasets, while advanced methods, like clustering algorithms, distance measurements, and machine learning models, have exhibited increased accuracy; however, each of these methods is prone to weaknesses in robustness when used independently. Despite improvements, many of these current approaches are either inadequate for dealing with highly nonlinear and noisy wind data, are computationally expensive, or are based on certain assumptions.

In response to these limitations, however, interest has been growing in hybrid models that combine multiple techniques to their potential to combine the best aspects of different approaches. Outlier detection and cleaning is the mechanism hybrid models can excel at by leveraging clustering for grouping data, distance metrics for precise anomaly detection, and machine learning for non-linear pattern recognition because they can combine the performance of multiple techniques. However, the literature does not provide comprehensive frameworks for properly integrating these components for wind data preparation.

This paper contributes a novel hybrid model that integrates Fuzzy C-Means (FCM) clustering, Mahalanobis distance, and Artificial Neural Networks (ANN) to overcome the aforementioned flaws; this paper proposes a methodology that can be divided into four steps: Step 1: Pre-cleaning; to eliminate gross errors such as negative values. Step 2 (Clustering Initialization by FCM): The process begins by clustering the SCADA (Supervisory Control and Data Acquisition) Data by setting up the membership matrix and cluster centres before updating the membership values and cluster centres until we reach a point where everything stabilizes and converges correctly. Next, in Step 3 regarding Outlier Detection using Mahalanobis Distance, after the clustering is done, for each data point, it is important to calculate the Mahalanobis distance, where data points whose distance exceeds a certain threshold are marked as outliers. Step 4 involves refinement using ANN. Following the detection of outliers through Mahalanobis distance calculation, we then take the filtered data (excluding any outliers found earlier in the process) and use it to train an ANN. The artificial neural network (ANN) learns the regular patterns in the data it processes. The system improves in recognizing irregularities by identifying data points that differ from patterns it has acquired gradually. This method is essential for detecting and excluding any additional outliers that may have been overlook during the Mahalanobis distance phase. This combination ensures a robust outlier detection process, strengthening the reliability and coherence of wind data for subsequent uses. By address deficiencies in current methods this study offers a feasible solution to enhance the accuracy of wind datasets and promote their utilization in energy meteorology and environmental science fields.

In this paper, deleted outliers are confined to deleting negative points and real measurement errors and do not include actual data points that are scarcely observed because the actual data points represent important conditions that should be taken into consideration by wind producers to make adequate decisions in electricity markets.

1.2. Related work

Various outlier detection and cleaning modern AI techniques have been developed to address these issues. In [3] presents an IAO LSTM model that uses an Isolation Forest for filter outliers, synchronous squeeze wavelet transforms for cleaning noise, and an Aquila Optimizer-tuned LSTM to give better short-term power predictions for new wind turbines. In [4], it introduces C-LSTM, using an adaptive wind speed fixing method within LSTM to change forecasted wind speeds using past data, boosting short-term wind power guess accuracy across 25 turbines [5]. unveils CapSA-RVFL, which leverages the Capuchin Search Algorithm to best tune a Random Vector Functional Link network for wind power guessing, getting better RMSE MAE MAPE and R^2 results than standard RVFL kinds on four French turbine sets [6]. This paper uses a Relevance Vector Machine (RVM) improved with Improved Manta-Ray Foraging Optimization (IMRFO) to estimate monthly pan evaporation. In [7], Proposes a hybrid deep-learning model that merges Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks to forecast stock prices at the National Stock Exchange of India. The hybrid LSTM–CNN model surpass individual LSTM and CNN designs by decrease (RMSE) by around 15 %, especially excel in times of significant market turbulence. In [8], this research presents a technique that merges Bayesian change point detection with the quartile algorithm to detect and remove abnormal data points in the wind turbine power curve with real-world 10-minute wind power monitoring data sets—showcasing its practicality and impact on enhancing model precision. In [9], This study suggests a technique that effectively merges the quartile method with Random Sample Consensus (RANSAC) regression to filter out wind speed power data points. It is beneficial for dealing with the amounts of clustered data points and improving the precision of operational data for wind turbine generators.

1.3. Objectives and Contributions

The main objective of this research is develop and assess a hybrid model that merges Fuzzy C-Means (FCM) clustering, Mahalanobis distance, and an Artificial Neural Network (ANN) to detect and eliminate outliers in raw wind turbine data efficiently. This method aims to overcome the constraints of other traditional hybrid models and individual techniques by blending unsupervised clustering, statistical distance measures, and machine learning to improve data cleansing accuracy.

The novel contributions of this study is outlined as follows:

1. A framework consists of two-phase for outlier detection that utilize a combination of FCM clustering and Mahalanobis distance for initial filtering then employ an ANN to further refine anomaly.
2. Enhanced identification precision of structure and statistical anomaly in wind turbine SCADA data, resulting leading to cleaner power curve modelling and more reliable forecasting data.
3. Validation of practical efficiency through testing on actual SCADA dataset, verifying model's strength in managing noise and erratic operational data.
4. Introducing a Combined Accuracy (CA) index as comprehensive measure for assess performance combining RMSE MAPE and R^2 into one understandable metric.
5. Simplifying computational method and enhancing practicality by bypassing decomposition methods ensure ease of deployment while upholding model simplicity and performance.

The integrated method provides a scalable and effective solution for preparing wind energy data with clear benefits for enhance turbine performance analysis, detecting anomalies, and assessing wind resources.

The paper's organisation is as follows: Section 1 is the introduction. Section 2 is the wind turbine power curve. In section 3, Some popular power curve models are introduced briefly. Section 4 presents the methodology related to the proposed hybrid model and the proposed strategy for power curve modelling. Section 5 presents the experimental results. The discussion is introduced in Section 6. Finally, section 7 concludes the whole paper.

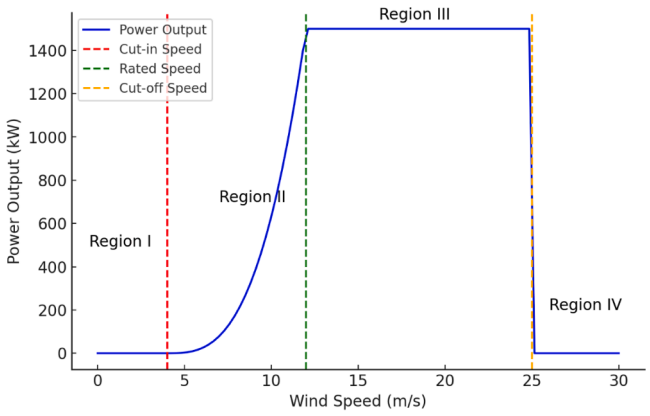

Fig. 1. Ideal wind turbine power curve.

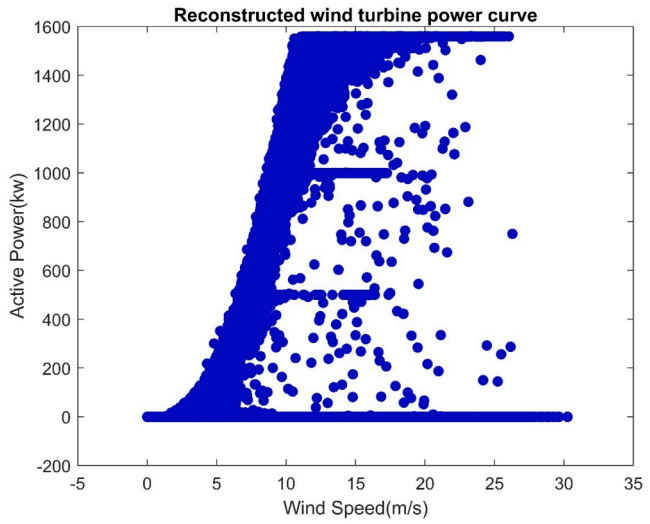

Fig. 2. Reconstructed wind turbine power curve.

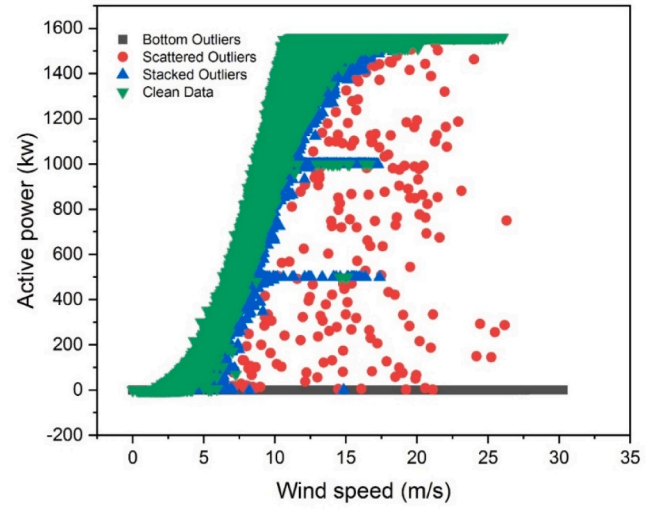

Fig. 3. Distinguish between Outliers using DBSCAN Clustering in WTPC.

2. Wind turbine power curve

The wind turbine power curve is calculated using data from the SCADA system, representing the relationship between active power and wind speed. It is a direct function of the wind turbine control system. WTPC abnormalities are caused by overdating, pitch malfunction, pitch controller malfunction, wind speed underreading, dirt, bugs, or icing on blades and down rating, among others [10].

As in Fig. (1), the ideal wind turbine power curve is a standardized representation of a turbine's performance derived from controlled conditions. It depicts the wind speed versus power output relationship, assuming typical conditions such as smooth airflow and minimal turbulence. It is Static and free from operational irregularities and outliers. When the wind speed in Region I is below a cut-in speed or minimum threshold, no power is produced. A rapid increase in power is produced in Region II, which is between the cut-in and the rated speed. In Region III, rated output remains constant until the cut-off speed is reached. Beyond this speed (Region IV), the turbine is turned off to protect its inner parts from strong winds. It means that no power is produced in this region.

On the other hand, the wind turbine power curve in Fig. (2) is reconstructed from the operational data taken from the turbine at its operating location. It is dynamic and represents actual performance, incorporating the effects of turbulence, terrain, and other operational deficiencies based on the turbine's actual location. During preprocessing anomaly and outlier is detected and eliminated. The key difference lies in the context: The ideal wind turbine power curve is theoretical and idealized, while the reconstructed curve represents practical performance in the field.

The power characteristics of the rotor, generator, gearbox ratio, and efficiencies of various components can also be used to predict the rough form of the power curve for a particular machine. Due to the transfer functions of the generators available, the conversion of wind energy into actual power is non-linear [11].

Theoretical wind turbine rotor power (P) is:

$$P = 0.5\rho\pi R^2 C_p u^3 \quad (1)$$

Where u is the wind speed, C_p is the power coefficient, R is the rotor's radius, and ρ is air density [12].

In the raw wind data, there are many unusual data points from different factors such as unexpected maintenance, wind turbine malfunctions, suspension directives, communication issues, electromagnetic interference, and severe weather conditions. These outliers show varying distribution patterns based on their distinct causes and can be categorized into three types: bottom-curve stacked outliers scattered outliers and stacked outliers, as shown in Fig. 3)

1. Bottom-Curve outliers or negative points (black): These points show almost zero power values at the curve's lower and may be due to issues like turbine damage, sensor faults, or suboptimal plant operation.
2. Scattered outliers(red): These points is spread out around the trend line and may be identified as anomalies due to unexpected events like severe weather, sensor malfunctions, signal noise etc. They typically are temporary and vary in characteristics.
3. Stacked outliers(blue): Horizontal groupings of points show equal power output under varied wind speeds, cause by communication or suspension problems.

Distinguish between outlier types according to the procedure cleared in [13], but using a DBSCAN algorithm.

The unsupervised machine learning algorithm of DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering and anomaly detection algorithm based on data point density. It classifies points as core, border, or noise. It defines core points as those with at least a minimum number of neighbours (MinPts) within a radius (ε) and border points as those with fewer but within a core point's radius. Moreover, the noise points do not meet these requirements. The process starts with a random point; when it is a core point, it grows to create a

cluster, and noise points otherwise remain unclustered.

In Fig. 3, green represents clean data (dense valid clusters), red dots are scattered outliers (noise points), blue indicates stacked outliers (small dense anomaly clusters), and the grey line is outliers at the bottom (low-density regions, possibly harmful measurements). DBSCAN, with its unique ability to handle non-linear data sets and its independence from predefined cluster numbers, is a powerful tool for analysing the wind turbine power curve.

After a meticulous pre-cleaning process, which includes deleting the bottom curve stacked curve, the scattered and stacked outliers are marked based on the DBSCAN algorithm.

Finally, the raw wind power undergoes a comprehensive cleaning process to ensure the accuracy of the wind power curve.

3. Wind turbine models

The performance of wind turbines can be represented by using the turbine power curve concept or utilizing basic equations of power derived from wind. The approach explained in [14] is known for producing imprecise and complex outcomes. Models based on the power curve concept usually get classified into parametric and nonparametric categories.

Nonparametric methods outperforming parametric ones [28]; various nonparametric strategies including K-Means, FCM (Fuzzy C-Means), Subtractive Clustering, SVR (Support Vector Regression), ANFIS (Adaptive Neuro-Fuzzy Inference System), and ANN (Artificial Neural Networks) were analyzed in this research to highlight their advantages and disadvantages among different wind conditions. The objective is to determine the most precise technique for representing the power curve reliable despite outliers which is crucial for integrating into our hybrid model later on.

Using nonparametric methods, the following assumption is resolved:

$$\mathrm{P} = \mathrm{f(u)} \tag{2}$$

For this purpose, nonparametric techniques are instrumental when they do not assume a specific functional form of the relationship. These methods build the curve from the observed data and are, therefore, very flexible and suitable for modelling the nonlinear nature of wind turbine power curves.

3.1. K-Means Clustering

K-means is one of the most popular techniques for grouping data into (K) groups of non-overlapping clusters based on similar data points. First, data preparation is done, including data collection and preprocessing to eliminate gross errors such as negative values. The algorithm is started by choosing the number of clusters (K), which can be through experience or using the elbow method or silhouette analysis, and then at random setting the (K) centroids. Each data point is assigned to the centroid to which it is closest, which is usually calculated as the distance between the point and each centroid in the assignment step [15]

In the update step, recompute the centroids as the mean of all data points assigned to each cluster. These steps of assignment and update are iteratively repeated until the centroids stabilize or a predefined number of iterations is reached, marking convergence. At this phase, the method has converged, and the final clusters represent segments of the power curve and help identify outliers as points that do not align well with any cluster. K selection is critical, and careful balance is needed to balance simplicity versus complexity since too few clusters can simplify the model too much. At the same time, too many can result in overfitting.

3.2. Fuzzy C-means clustering

In 1981, Bezdek proposed FCM for the first time [16]. FCM is a soft clustering algorithm, i.e., each point can belong to one or more clusters to some extent. It is beneficial for wind turbine power curve modelling. First, data collection and preprocessing of wind speed and power output are done. The clustering process involves choosing the number of clusters (C) and assigning a random initial membership value to each data point. Then, the membership values are iteratively updated based on how much each data point and centroid are similar. The centroids are then recalculated as a weighted average of the data points depending on their membership values. It is repeated until convergence; convergence is usually determined by some threshold for changes in the membership matrix to model the power curve accurately, taking into account data uncertainties and overlaps.

3.3. Subtractive clustering

Subtractive clustering identifies clusters in the data set by concentrating on high data point density regions. The high-density regions are expected behaviour, while low-density regions may indicate outliers.

The following is a description of the subtractive clustering algorithm:

Consider a set of n data points, $x_1, x_2,\ldots, x_n$, where x_i is a vector in the feature space. Without sacrificing generality, we assume that the feature space has been normalized to contain all data within a unit hypercube. Each data point is regarded as a possible cluster centre, and its ability to serve as a cluster centre is quantified. The potential of x_i is calculated using Eq. 3, which is denoted by P_i .

$$P_i = \sum_{j=1}^{n} exp(-\frac{\| x_{i-}x_j \|^2}{(r_a/2)^2} \tag{3}$$

Where $||\ ||$ indicates the Euclidean distance and $r_a$is a constant defining a neighborhood radius, r_a is a positive constant. A data point with many neighbouring points has a high potential value, whereas points outside r_a have minimal effect.

The first cluster centre c_1, is selected as the highest-potential point. The potential of c_1 is denoted as PotVal (c_1). The potential of each data point x_i is then reevaluated as follows:

$$P_i = P_i - \mathrm{PotVal}\ (c_1)\exp(-\frac{\| x_i - c_1 \|^2}{(r_b/2)^2} \tag{4}$$

$r_b = 1.5r_a$ is frequently employed to prevent cluster centres from being too close together. The data points close to the first cluster center will have lower potential and will unlikely to be selected as the next cluster centre. After reducing the potentials of all data points using Eq. 4, the data point with the most significant potential is selected as the second cluster centre. The potential of the remaining points is subsequently reduced once more. After identifying the c_k of the kth cluster, the potential is often modified as follows:

$$P_i = P_i - \mathrm{PotVal}(c_k)\exp(-\frac{\| x_i - c_k\|^2}{(r_b/2)^2} \tag{5}$$

Where c_k is the location of the kth cluster centre and PotVal (c_k) is its potential value.

Two conclusions can be drawn from the clustering procedure:

(a) A point with high potential has a greater chance of being chosen as the cluster's centre than one with low potential. Each cluster centre is an up-and-coming locale.

(b) Cluster centres are merely selected from the data points, regardless of whether the actual cluster centres exist in the dataset. In contrast, cluster centres are not necessarily located at a data point.

The following are the steps of the weighted mean subtractive clustering algorithm:

Step 1: Using Eq. 3, compute the probability of each data point; set the number of cluster centres to k = 1.

Step 2: Select the data point with the highest probability, denoted as c_k, and the data points surrounding c_k with a radius less than r_a, denoted as ($x_1^{(k)}$, $x_2^{(k)}$,…, $x_{m(k)}^{(k)}$). Then, the weighted mean cluster centre c_k^- is

computed as follows:

$$c_k^- = \frac{\sum_{j=1}^{m(k)} PotVal(x_j^{(k)} * x_j^{(k)}}{\sum_{j=1}^{m(k)} PotVal\left(x_j^{(k)}\right)} \quad (6)$$

Where $m(k)$ is the number of data points surrounding c_k with a radius smaller than r_a.

Step 3: The potential of each data point is changed as follows:

$$P_i^{(k+1)} = P_i^k - PotVal(c_k^-) \exp\left(\frac{\| x_i - c_k^- \|^2}{(r_b/2)^2}\right) \quad (7)$$

Where:

$$\text{PotVal}(c_k^-) = \sum_{i=1}^{n} \exp\left(-\frac{\| c_k^- - x_i \|^2}{(r_a/2)^2}\right) - \sum_{j=1}^{k=1} \exp\left(-\frac{\| c_k^- - c_j^- \|^2}{(r_b/2)^2}\right) \quad (8)$$

Step 4: As long as the "stop" criteria are met, then halt the process; otherwise proceed to Step 2 and set k = (k + 1).

In weighted mean subtractive clustering, the cluster centre is not defined by a single data point but by all surrounding data points. The high-potential point has an outsized influence on the cluster's core. The impact is measured by potential; the more significant the potential, the greater the effect.

3.4. Support Vector Regression (SVR)

SVR attempts to find a function *f(x)* such that for all training data, its output deviates from the actual target values y_i by no more than ϵ, and as otherwise as possible, this function *f(x)* should be as flat as possible.

It is done by solving the following optimization problem:

$$\min_{w,b,\xi,\xi^*} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} (\xi_i + \xi^*_i) \quad (9)$$

Subject to

$$y_i - \langle w, \phi(x_i) \rangle - b \leq \epsilon + \xi_i \quad (10)$$

$$\langle w, \phi(x_i) \rangle + b - y_i \leq \epsilon + \xi^*_i \quad (11)$$

$\xi_i, \xi^*_i \geq 0$, $i = 1, \ldots\ldots, n$where:

- w is the weight vector,
- b is the bias term,
- ξ_i and ξ^*_i are slack variables representing the degree of deviation beyond ϵ
- C is a regularization parameter that determines the trade-off between the model's complexity and the amount up to which deviations more significant than ϵ are tolerated,
- $\phi(x_i)$ is a nonlinear function that maps the input space into a higher-dimensional feature space.

In the context of the wind turbine power curves, the application of SVR for outlier detection can be demonstrated as follows: First, sum the wind speed and power output data. Then, train the SVR model on this data to learn the typical link between wind speed and power output. After the model is trained, it is utilized to predict power output for given wind speeds and residuals. The difference between actual and predicted outputs—is calculated. Finally, a threshold for these residuals is set, and any data points with residuals further than this threshold are marked as outliers, indicating some deviation from typical turbine performance [17,18]

3.5. Artificial neural network (ANN)

a. Modeling the Power Curve with ANN:

An ANN can be trained to predict the expected power output (P_{pred}) for given input features such as wind speed (v), air density (ρ), and other environmental factors. The general form of the ANN model can be expressed as:

$$P_{pred} = \text{ANN}\ (v,\ \rho,\ \text{other factors}) \quad (12)$$

Artificial neural network (ANN) trains to learn power generation pattern of turbine by minimizing difference between expected and actual power outputs.

b. Outlier Detection:

After successfully training an artificial neural network (ANN) the next phase focus on using it to detect anomalies by compare its predictions with the observed power output in practice (P_{obs}). A calculation is performed for each observation to determine the residual (r).

$$r = P_{obs} - P_{pred} \quad (13)$$

Statistical thresholds can be set to detect deviations in data analysis scenarios; for instance, when the residual value surpasses many standard deviations (σ) from the average remainder value, that data point may be marked as an outlier.

$$|r| > k \times \sigma \quad (14)$$

where k is a chosen threshold value (commonly 2 or 3).

This paper utilizes the Levenberg-Marquardt algorithm (LMA) [19, 20]. It combines the advantages of the convergent steepest descent technique and Newton's approach, which is often fast around the optimal solution.

3.6. ANFIS (Adaptive neuro fuzzy inference system)

ANFIS has the advantages of a neural network, a fuzzy control system and the ability to learn automatically. ANFIS was created as a nonlinear function model with fuzzy rules and membership function parameters that can be adjusted during the training stage [21,22]. Sugeno and Mamdani are two forms of fuzzy inferences differing in consequence of the set of fuzzy rules and defuzzification processes used.

ANFIS with Sugeno inference has two inputs (x,y) and fuzzy sets of (A, B, C, D) if-then rules, which are stated as follows:

If x is A and y is C then $f_1 = r_1 + P_1 x + q_1 y$ (15)

If x is B and y is D then $f_2 = r_2 + P_2 x + q_2 y$ (16)

4. Methodology

4.1. Dataset description

The proposed hybrid model was assessed using data from a wind farm, which was selected as a case study in this paper, namely an onshore wind farm in Khorasan, northeast Iran. The wind turbine that is to be considered is the regulated-pitch type 1.5MW WD77, with a height of the hub of 60 m, a cut-in speed rate of 3 m/s, a rated speed of 11 m/s, and a cut-out speed rate of 25 m/s. Wind speed, real power, and wind direction were measured, and 5-minute data for 330 days were gathered, i.e., 94,271 samples. The wind speed (m/s) was measured at the mast. The data set was collected in four seasons: from 9 April. 2013, at 15:05:00 P.M. to 4 March 2014, at 23:55:00 P.M.

The data of this wind farm contains types of disturbances and irregularities that can be attributed to sensor inaccuracies or failures and various factors, like blade issues or maintenance problems, as well as fluctuations in turbine performance due to environmental conditions such as low wind speeds or incorrect pitch angles being set up wrongly on the wind turbine itself. This makes it a relevant and challenging dataset to assess anomaly detection techniques, which test the

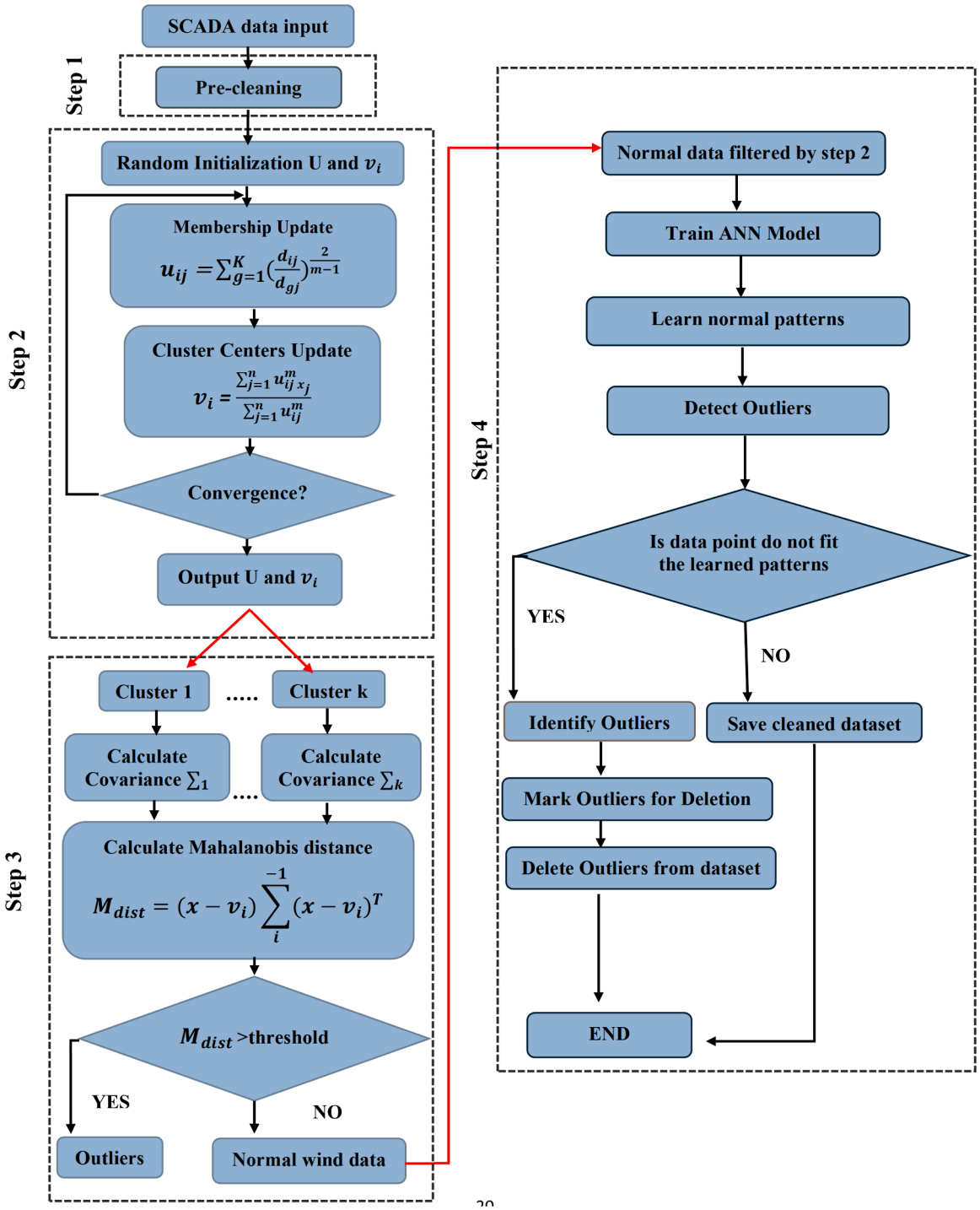

Fig. 4. Flowchart of the proposed hybrid model for detecting and deleting outliers.

robustness of any cleaning method. The characteristics of this dataset provide an ideal testbed for evaluating the layered hybrid model we propose. This case study significantly affects the results by exposing the model to a wide range of real-world anomaly conditions, thus demonstrating its practical applicability and generalizability. Data filtering becomes crucial to ensure that a reliable and precise wind turbine model is established without any hindrances from the interferences.

The exclusion of decomposition methods from the outlier detection process in this study was deliberate, and this decision is considered suitable from both technical and contextual standpoint. The wind speed data analysed in the study were obtained from sensors mounted on masts, ensuring the provision of cleaner and more stable signals that reduce the impact of high-frequency noise. Additionally, the SCADA data provided by the wind farm had already undergone standard pre-processing and filtering routines, reducing the need for additional signal decomposition. More importantly, the proposed hybrid model (FCM + Mahalanobis distance + ANN) aims to detect structural or statistical outliers—those that deviate significantly from the underlying patterns or cluster centres—not to capture subtle fluctuations in the signal. The slight fluctuations that are commonly focused on in decomposition methods have minimal effects on identifying true anomalies which typically appear as significant deviations.

Table 1
Shows performance evaluation metrics.

Indicator	Description
$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	Smaller is better.
$\mathbf{MAPE} = \frac{100}{n}\sum_{i=1}^{n}\lvert\frac{y_i - \hat{y}_i}{y_i}\rvert$	Smaller is better.
$\mathbf{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Smaller is better.
$R^2 = 1\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - y^-)^2}$	R^2 closer to 1, Perfect fit
$\mathbf{CA} = \frac{1}{3}\left(\frac{RMSE_{min}}{RMSE} + \frac{MAPE_{min}}{MAPE} + \frac{R^2}{R^2_{max}}\right)$	Higher CA values (closer to 1) indicate better overall accuracy.

Table 2
Performance of proposed hybrid cleaning model.

	Noise level (Std Dev)	Accuracy vs Noise	System Dynamics	Accuracy vs Dynamics
1	0.01	0.98		
2	0.05	0.95		
3	0.1	0.9		
4	0.2	0.85		
5			Static	0.96
6			Moderate	0.92
7			Rapid	0.88

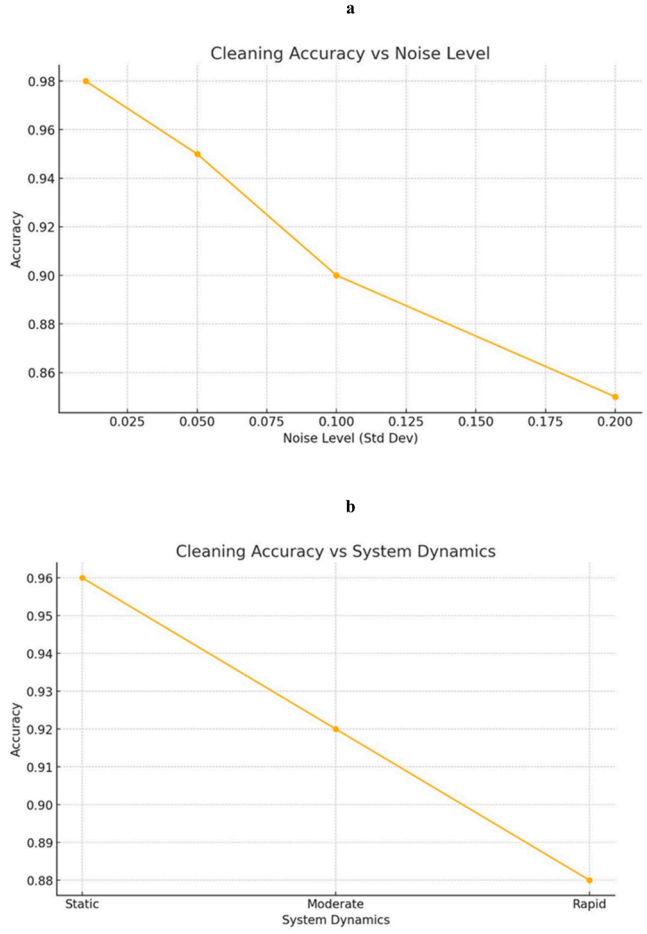

Fig. 5. Shows behavior of the proposed hybrid model under a) varying noise vs. cleaning accuracy and b) dynamic scenarios vs. cleaning accuracy.

Also by avoiding decomposition method it lessen the computational burden streamlining the model and enhances processing speed. It is particularly beneficial when handling large datasets or seeking possible real-time applications. While the decomposition method might have its benefits in certain situations, leaving it out in this instance does not undermine the model effectiveness. On the contrary, it supports a more efficient and practical structure.

4.2. Proposed hybrid model design

Breakdown of the methodology into four main steps, as follows:
Step 1 (Pre-cleaning)
Data pre-cleaning aims to clean the bottom-curve outliers or negative points, as shown in Fig. 3. The bottom-curve outliers simultaneously satisfy the conditions of (17). Therefore, such outliers can be cleaned according to (17).

$$\begin{cases} V > v_c \\ P < 0 \end{cases} \tag{17}$$

Step 2 (Clustering Initialization by FCM):
The process commences with clustering the SCADA (Supervisory Control and Data Acquisition) data. It includes initialization of the membership matrix and cluster centres and iterative updating of the membership values and cluster centres until convergence is achieved.
Step 3 (Identifying Abnormal Points Using Mahalanobis Distance):
After the clustering process, each data point is evaluated based on the Mahalanobis distance within its respective cluster to validate the existence of outliers. This distance metric helps identify data points far from the cluster centre and labels them as outliers if the distance exceeds a certain threshold. Real measurement errors identified in this step as outliers often exhibit extreme deviations from expected values, like unusually high or low power outputs compared to the usual pattern observed. The Mahalanobis distance can be calculated using the formula below.

$$\boldsymbol{M_{dist}} = (\boldsymbol{x} - \boldsymbol{\mu})\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})^T \tag{18}$$

Where x is the sample in the observed cluster μ, $\sum$ is the corresponding cluster centre and covariance of all samples in the same cluster.
Step 4 (Refinement using ANN):
In Step 4, the ANN is utilized as a refining technique to find any remaining outliers that might not have been caught by the Mahalanobis distance-based method. Using clustering in conjunction with Mahalanobis distance and the ANN's ability to identify anomalies based on learned characteristics ensures a nearly exhaustive search for outliers in the dataset. Using this four-step strategy increases the accuracy of data cleaning and enhances the credibility of the analysis of wind turbine power curves.

4.3. Parameters selection

This study pays close attention to selecting important parameters to enhance strength and precision of the suggested hybrid model to identify outliers.

1. Optimal Number of Clusters – Fuzzy C-Means Clustering

The elbow method was used to choose the optimal number of clusters in the fuzzy C-means clustering process. This method calculates the sum of squared distances between data points and their respective cluster centres for each k value. As the value of k increases, this metric should decrease since a higher number of clusters aligns more accurately with data points. Then, graph the inertia against k values. The point where a noticeable drop in the rate of decrease occurs represents the "elbow," signifying that adding additional clusters does not notably enhance data

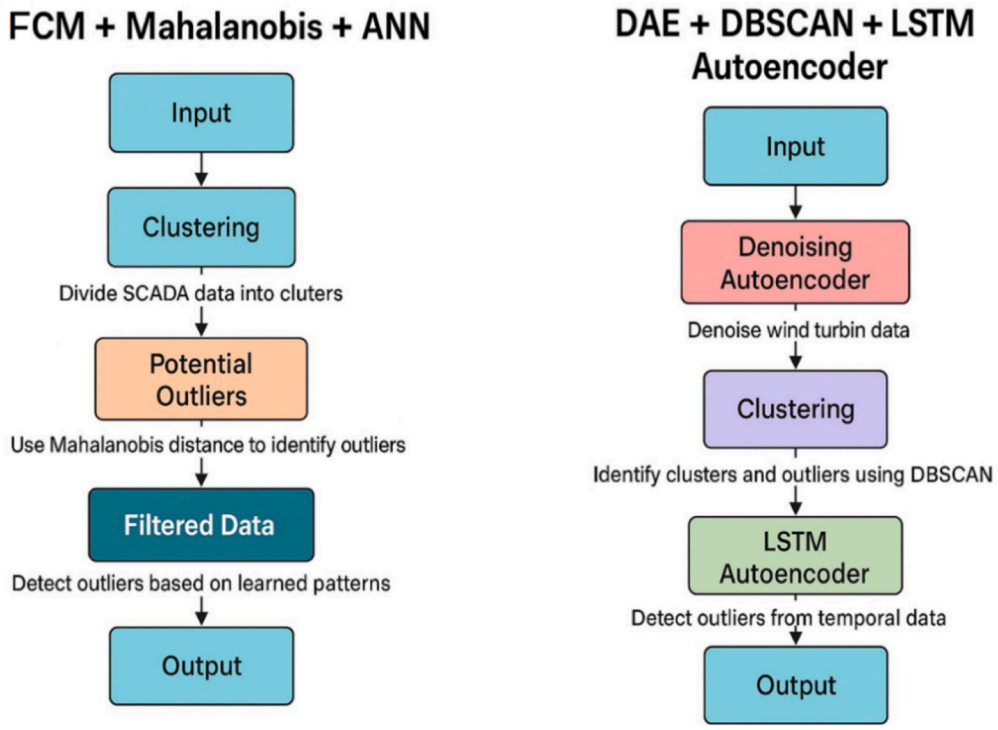

Fig. 6. Comparison between the proposed hybrid model and an advanced one.

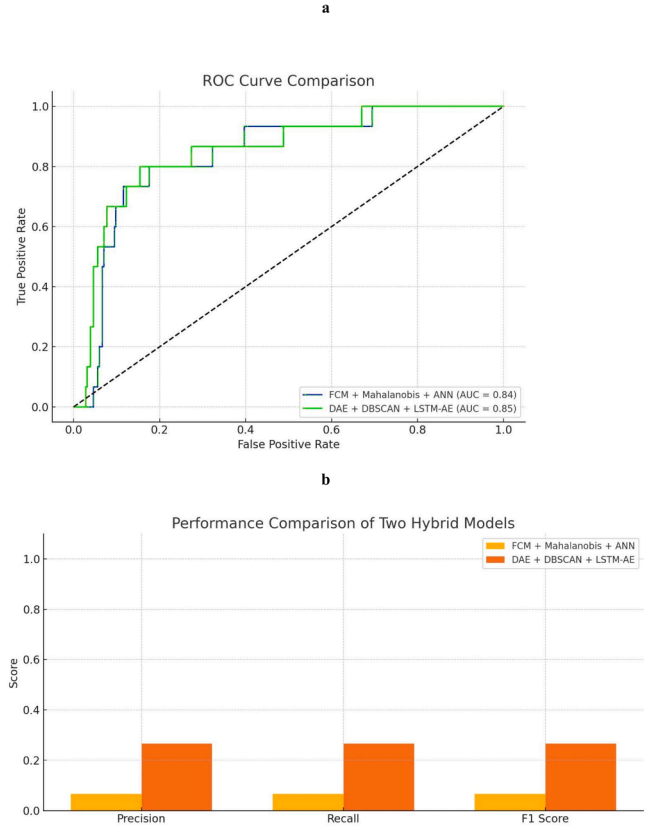

Fig. 7. Shows a comparison between the proposed hybrid model and an advanced one regarding a) AUC, b) precision, recall, and F1 score.

fitting. It is a good candidate for ideal k value. After testing a variety of clusters, we find that too few clusters result in oversimplification of the model while many number of clusters lead to overfitting.

2. Threshold Selection – Mahalanobis Distance

A fixed, statistically grounded threshold corresponding to the 95 % confidence level of the Chi-squared distribution was selected. It means that 95 % are confident that the observed behaviour is normal, the remaining 5 % is considered rare or abnormal; this threshold ensures that only data points exhibiting significant deviation from the multivariate norm are flagged as outliers. We tested the confidence level of the Chi-squared from 90 % to 99 %; we wanted to be more selective and catch only the most extreme deviation outliers. It is important to note that even if these are outliers, still the actual data points represent important conditions that should be taken into consideration by wind producers to make adequate decisions in electricity markets.

3. ANN Architecture – Hidden Layers and Number of Neurons

Two hidden layers in a network design were selected based on empirical evaluation. This configuration provided sufficient ability to capture complex, non-linear relationships in the data without introducing overfitting or excessive computational overhead. We have tested with more than two hidden layers and showed diminishing returns in performance, while simpler structures underperformed, particularly in detecting subtle residual anomalies. The candidate number of neurons can be selected from the set (20, 40, 60, . ., 200). The optimal number of neurons also changes slightly by season. Our experiment proves that 40 to 100 neurons are the optimal number.

4.4. Proposed hybrid model workflow

A four-step process describes the experiment to preprocess and analyze SCADA data for outlier cleaning. First, in step 1, the dataset is preprocessed to eliminate gross errors such as negative values. Then, In step 2, fuzzy C-means (FCM) clustering is applied to set up a membership

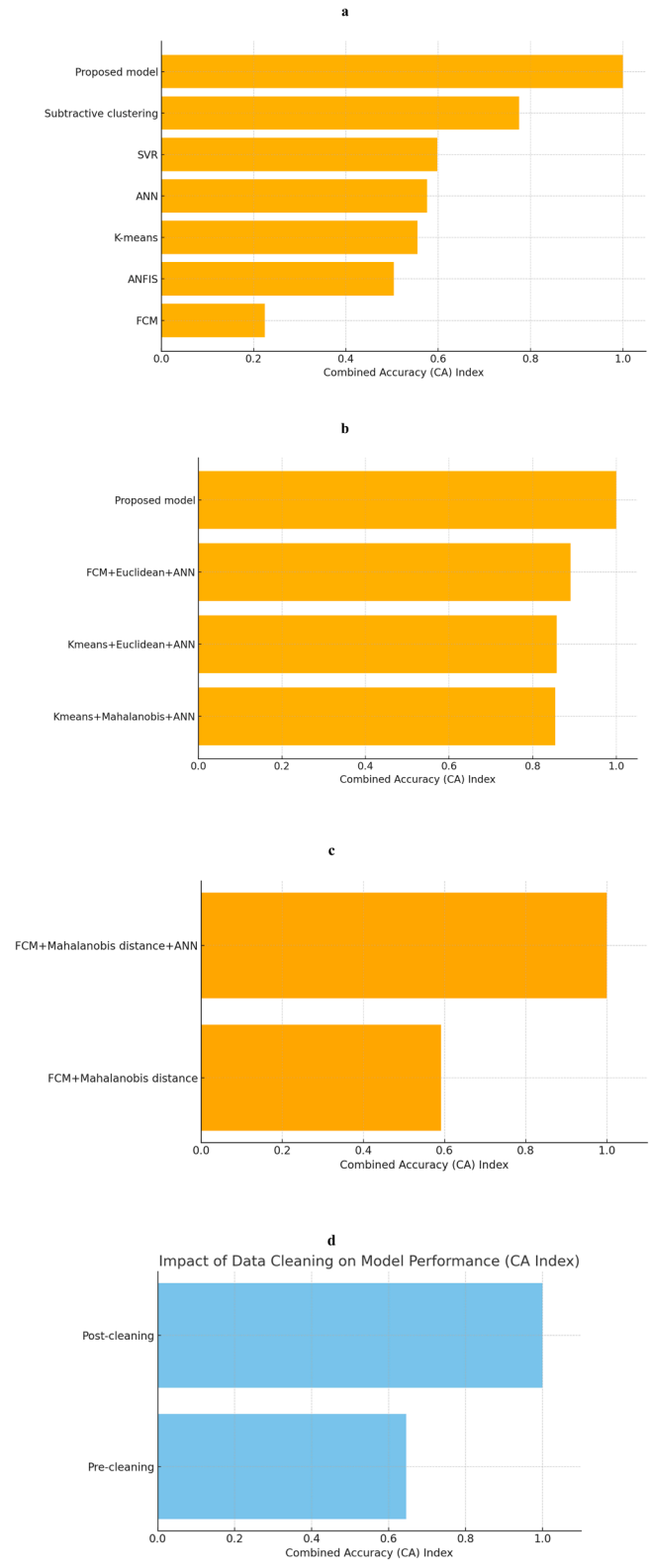

Fig. 8. Shows a), b), c), and d) proposed hybrid model comparison based on CA index.

matrix and cluster centers. Membership values and centres are iteratively updated until convergence. This process segments data into clusters to establish a framework for future investigation. Moving on to step 3, Mahalanobis distance is computed for each data point in the cluster. This metric accounts for multivariate relationships, flagging

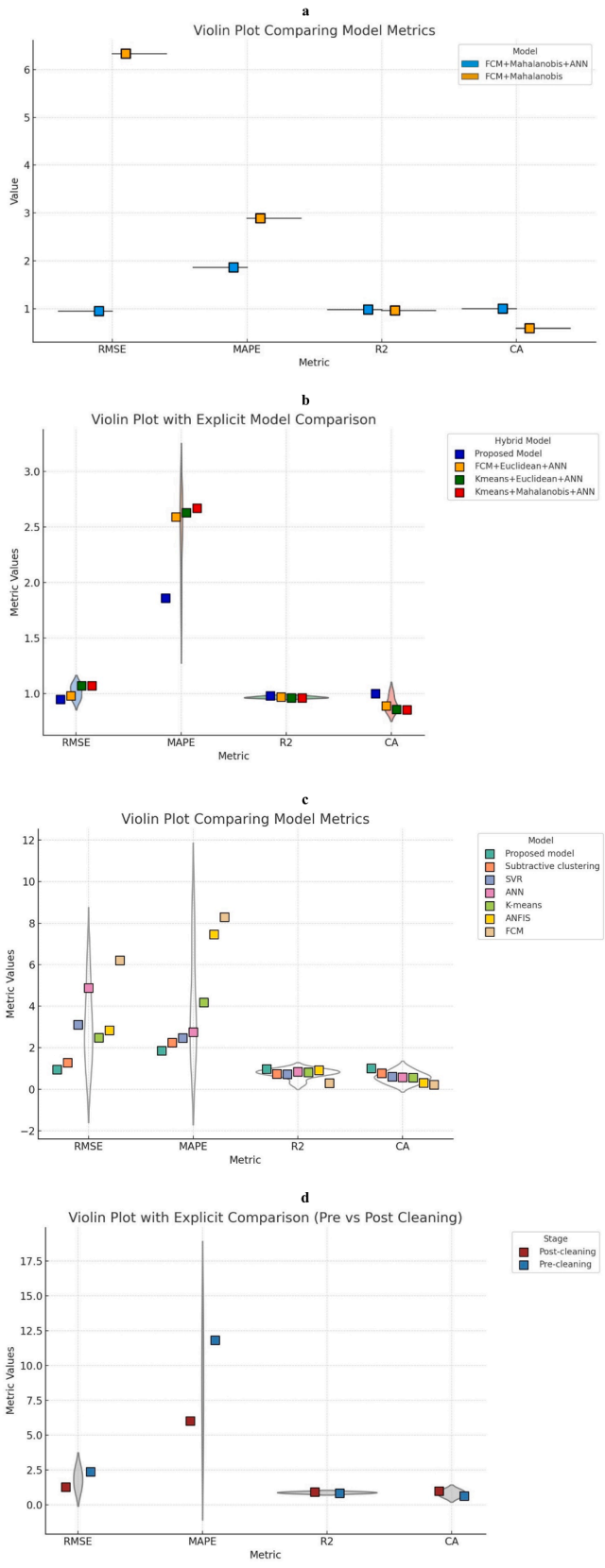

Fig. 9. shows a), b), c), and d) the violin plots to compare the results.

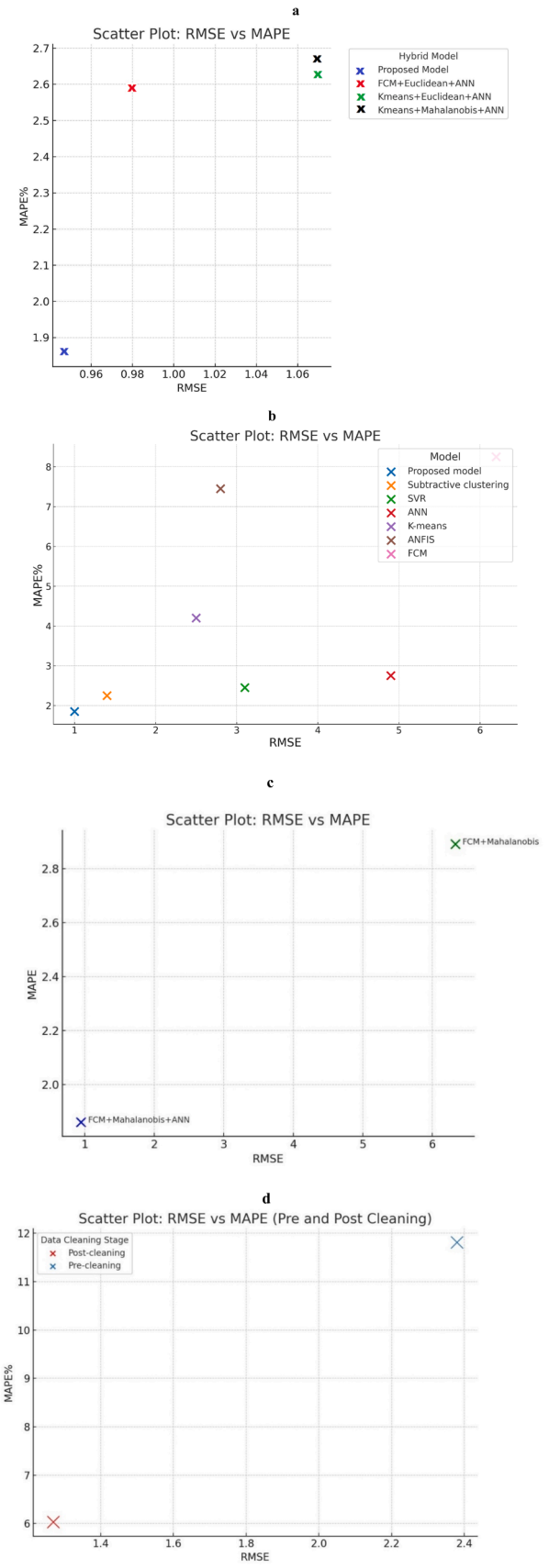

Fig. 10. shows a), b), c), and d) the scatter plots to compare the results.

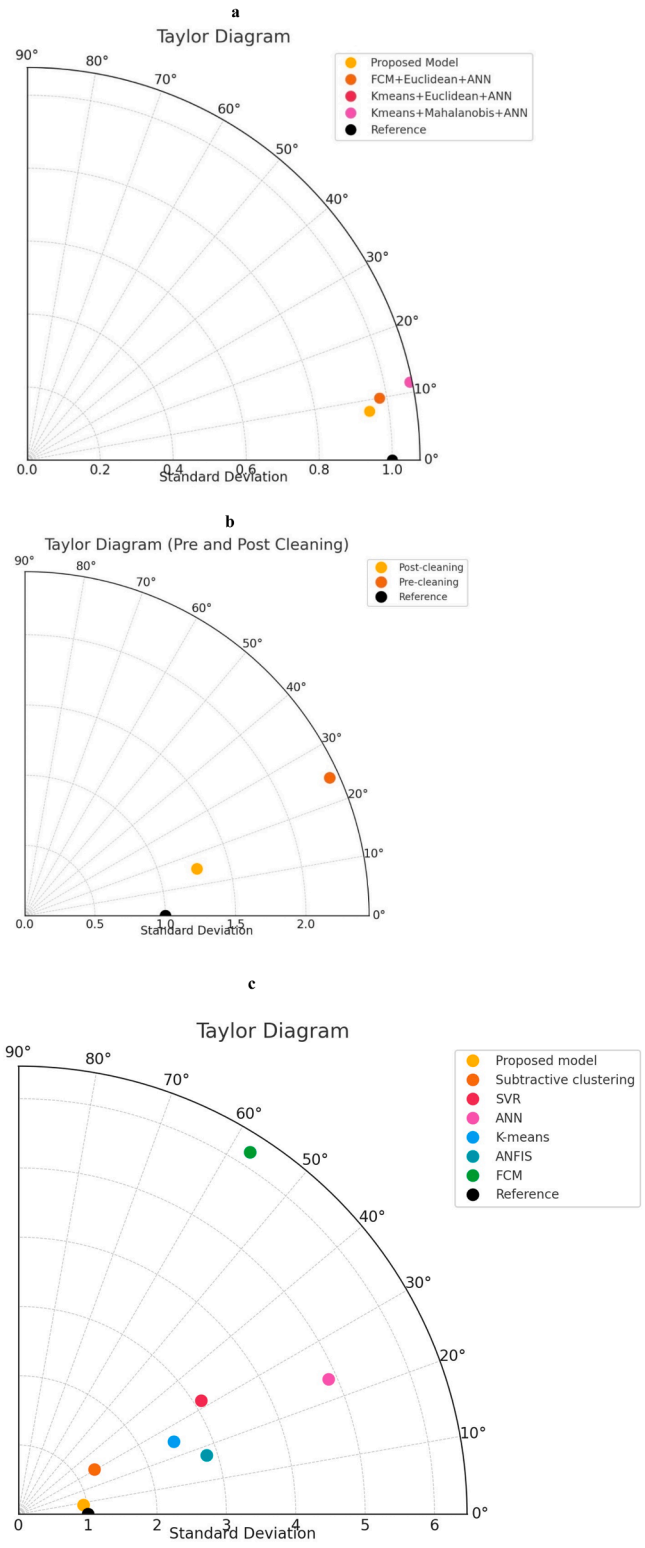

Fig. 11. shows a), b), and c) the Taylor plots to compare the results.

points above a specific threshold as potential outliers. using this distance measure because it is sensitive to multivariate relationships. In step 4, the process is refined by training an ANN on the clean dataset (i.e., without the initial outliers). The artificial neural network (ANN) learns normal data patterns and then can identify anomalies far from the normal patterns. This hybrid approach uses statistical and machine learning approaches for reliable outlier detection.

Table 3
Shows the performance analysis of algorithms used to detect outliers in the WTPC and clean it by dealing with data for each season separately.

Model	Winter Dec 1 – Feb 28/29		Spring March 1 - May 31		Summer June 1 - August 31		Autumn Sep 1 - November 30	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
K-Means	**3.077052**	**1.420594**	**3.333446**	**1.026946**	**6.160636**	**3.005412**	**3.361807**	**1.114035**
FCM	**6.679673**	**5.543114**	**3.986155**	**2.990391**	**11.243318**	**9.618561**	**5.639856**	**4.906998**
Subtractive	7.605348	5.922993	12.689163	11.591969	13.678660	12.686417	9.232497	7.269038
SVR	38.268404	5.805200	46.538115	7.734021	47.930697	14.705581	17.918441	5.435646
ANFIS	27.491470	18.838753	26.711999	20.188882	41.921586	28.579465	17.448433	13.819580
ANN	10.560908	15.853073	2.391203	142.268006	9.369598	148.745225	10.206433	15.720202

Table 4
Shows that adding the ANN improves the model's accuracy and effectiveness for detecting and removing outliers.

	Hybrid model	RMSE	MAPE %	R^2	CA
1	FCM+Mahalanobis distance+ANN	0.9475	1.86	0.98	1
2	FCM+Mahalanobis distance	6.3329	2.89	0.96	0.59

Table 5
Compares the proposed hybrid model with the traditional one by dealing with all the data.

	Hybrid model	Training function	RMSE	MAPE %	R^2	CA
1	Proposed model	LM	0.9475	1.86	0.98	1
2	FCM+Euclidean distance+ANN	LM	0.98	2.59	0.97	0.890
3	Kmeans+Euclidean distance+ANN	LM	1.07	2.63	0.96	0.857
4	Kmeans+Mahalanobis distance+ANN	LM	1.07	2.67	0.96	0.854

Table 6
Compares the proposed hybrid model with the single methods by dealing with all the data.

	Model	RMSE	MAPE %	R^2	CA
1	Proposed model	0.947	1.86	0.98	1
2	Subtractive clustering	1.272	2.25	0.74	0.775
3	SVR	3.103	2.47	0.72	0.597
4	ANN	4.882	2.75	0.84	0.575
5	K-means	2.474	4.18	0.82	0.554
6	ANFIS	2.843	7.46	0.91	0.503
7	FCM	6.207	8.29	0.29	0.224

Table 7
Shows a reduction in prediction error after detecting outliers and cleaning them.

	Stage	RMSE	MAPE %	R^2	CA
1	Post- cleaning	1.27	6.03	0.93	1
2	Pre-cleaning	2.38	11.81	0.83	0.645

4.5. Case study description and justification

The proposed hybrid model offers several advantages over traditional hybrid models or individual methods. By merging fuzzy C means clustering, Mahalanobis distance filtering and an ANN in a two-stage process, it not only detects both evident and subtle outliers but also enhance anomaly detection by learning patterns from pre-filtered data—significantly reducing inaccurate results. Training the ANN on already processed data further enhance its ability to identify regular operational patterns resulting in improved overall accuracy. When implemented on actual SCADA datasets, this leads to significantly enhanced turbine power curves, more dependable wind resource assessments, and performance analyses, as well as increased confidence in energy output predictions. In conclusion, this integrated approach enhances data quality and model efficiency downstream, establishing itself as a solid and reliable solution for wind energy analytics (Fig. 4).

5. Experimental results

5.1. Experimental setup

The software is configured with MATLAB R2020b, which uses the Fuzzy Logic Toolbox for FCM (Fuzzy C Means), the Statistics and Machine Learning Toolbox for Mahalanobis distance calculations, and the Deep Learning Toolbox for implementing Artificial Neural Networks (ANN). In addition to this setup in MATLAB R2020b, Python 3.8 is also employed along with libraries like scikit fuzzy for FCM clustering techniques; scipy is utilized for computing Mahalanobis distances; NumPy and Pandas are used for data manipulation and preprocessing tasks; TensorFlow or PyTorch are chosen options, for improving and training ANNs; finally, visualizations are done using Matplotlib and Seaborn libraries.

5.2. Evaluation metrics

The nonparametric and hybrid models are evaluated using four goodness-of-fit indexes: RMSE, MAPE, MAE, and R^2 .

The advantages of adopting these indexes over others to assess model performance comprehensively. RMSE is sensitive to significant errors while MAE offers a stable measure that isn't affected by outliers. Expressing errors as percentage helps enhance interpretability across various data scales with the use of MAPE. Additionally, R^2 shows how effectively the model captures variance in target variable. Collectively these metrics provide a thorough understanding of model accuracy error distribution and explanatory capacity. To strengthen evaluation, Combined Accuracy (CA) index was utilized. This index merges important metrics into a single score for easier and more coherent comparison of model performance. It simplifies understanding and aids in pinpointing well-balanced and dependable models particularly when there are trade-offs among various performance aspects. Including this aligns with current practices in model evaluation and boosts credibility of comparative analysis. They are expressed as follows.

Where $\widehat{y}_i$ are the predicted values and y_i are the observed values and y^- is the mean of observed values, and n is the number of samples [23] and $RMSE_{min}$ and $MAPE_{min}$ are the best (smallest) RMSE and MAPE among all models while R^2_{max} is the best (largest) R^2 among all models.

Table 1 illustrates performance evaluation metrics; lower RMSE, MAPE, and MAE values indicate better model performance. However, the closer the value of CA and R^2 are to 1, the better the model.

Table 8
Demonstrates how the model's performance varies with different numbers of neurons (from 20 to 200) by dealing with data for each season separately.

Proposed model	No. of Neurons	Winter			Spring			Summer			Autumn		
		MAPE %	RMSE	R^2	MAPE %	RMSE	R^2	MAPE %	RMSE	R^2	MAPE %	RMSE	R^2
FCM+ Mahalanobis distance + ANN	**20**	**0.79**	**0.25**	**1**	**0.67**	**0.57**	**0.98**	**1.07**	**0.36**	**1**	**1.97**	**0.74**	**0.98**
	40	**0.80**	**0.26**	**1**	**0.50**	**0.39**	**0.99**	**1.24**	**0.37**	**1**	**2.01**	**0.74**	**0.98**
	60	**0.97**	**0.25**	**1**	**0.67**	**0.51**	**0.98**	**1.09**	**0.38**	**0.99**	**2.10**	**0.63**	**0.98**
	80	**0.78**	**0.26**	**1**	**0.48**	**0.36**	**0.99**	**1.93**	**0.45**	**0.99**	**1.93**	**0.76**	**0.98**
	100	**0.93**	**0.24**	**1**	**0.40**	**0.35**	**0.99**	**1.41**	**0.38**	**0.99**	**1.91**	**0.75**	**0.98**
	120	**0.91**	**0.27**	**1**	**0.41**	**0.35**	**0.99**	**0.98**	**0.35**	**1**	**2.05**	**0.67**	**0.98**
	140	**0.79**	**0.26**	**1**	**0.51**	**0.40**	**0.99**	**1.32**	**0.36**	**1**	**1.90**	**0.75**	**0.98**
	160	**0.88**	**0.26**	**1**	**0.41**	**0.36**	**0.99**	**1.15**	**0.35**	**1**	**2.07**	**0.71**	**0.98**
	180	**0.79**	**0.27**	**1**	**0.47**	**0.36**	**0.99**	**1.18**	**0.36**	**1**	**1.92**	**0.76**	**0.98**
	200	**0.92**	**0.28**	**1**	**0.49**	**0.36**	**0.99**	**1.21**	**0.37**	**1**	**1.93**	**0.76**	0.98

Table 9
Compares the proposed hybrid model with the advanced one by dealing with all the data.

Model	RMSE	MAPE (%)	R^2	CA
Proposed model (FCM +Mahalanobis distance +ANN)	0.9475	1.86	0.980	1
Advanced model (DAE + DBSCAN + LSTM AE)	0.9320	1.82	0.981	0.99

5.3. Comparative analysis

5.3.1. Robustness of the proposed hybrid model under noise and varying dynamics

We will check the proposed hybrid model on different noise levels, such as small, medium, and big, and system dynamics (stationary, slowly changing, quickly changing). The results are shown in Table 2.

The proposed hybrid model keeps good accuracy under moderate changes and gets worse smoothly as noise or dynamics intensify:

- Noise robustness: Accuracy stays above 95 % for noise levels up to 0.05 std, falling to near 90 % at 0.1 std and 85 % at 0.2 std, see it in Table 2, and Fig. 5(a).

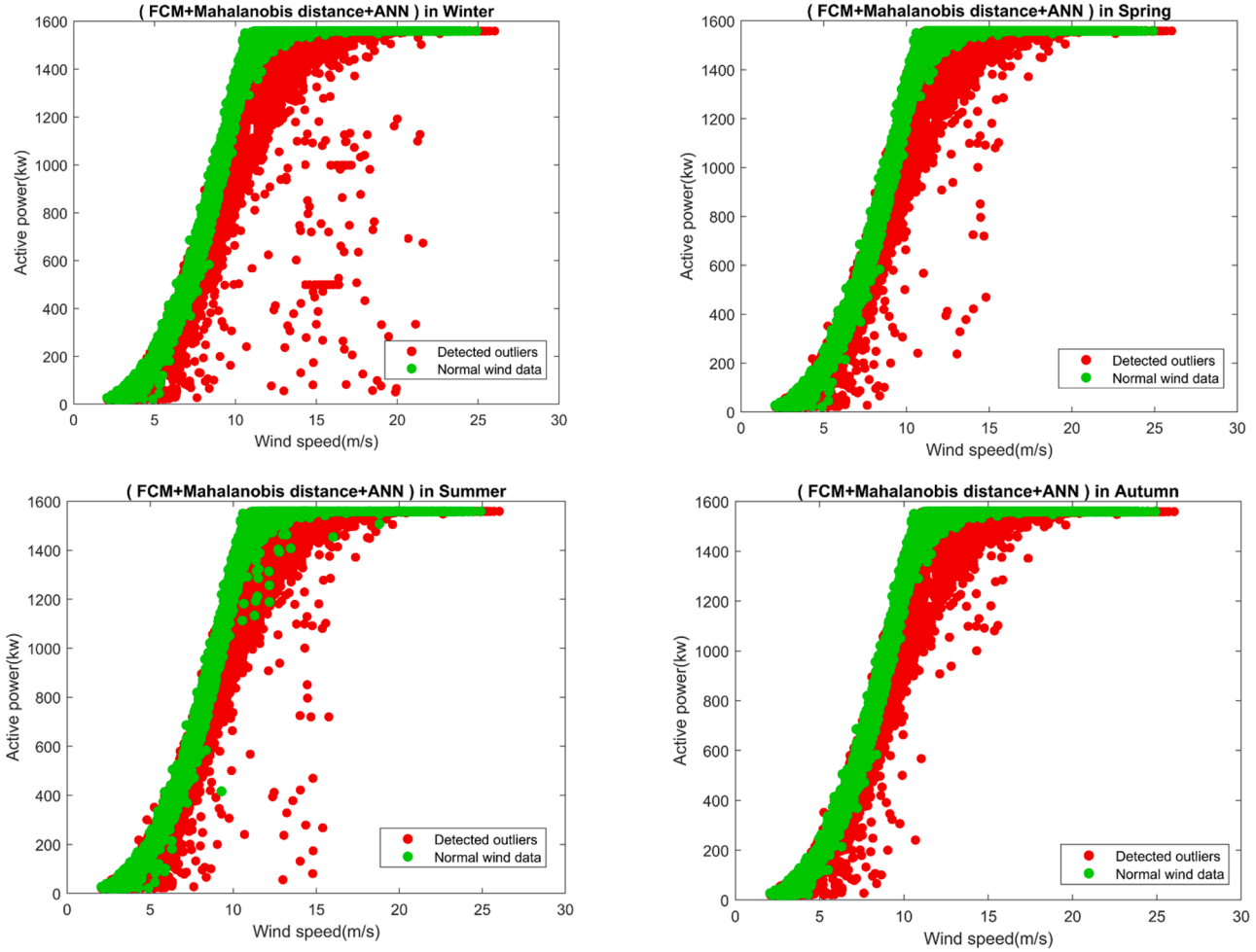

Fig. 12. Outliers detected by the proposed hybrid model FCM+ Mahalanobis distance +ANN by dealing with data for each season separately.

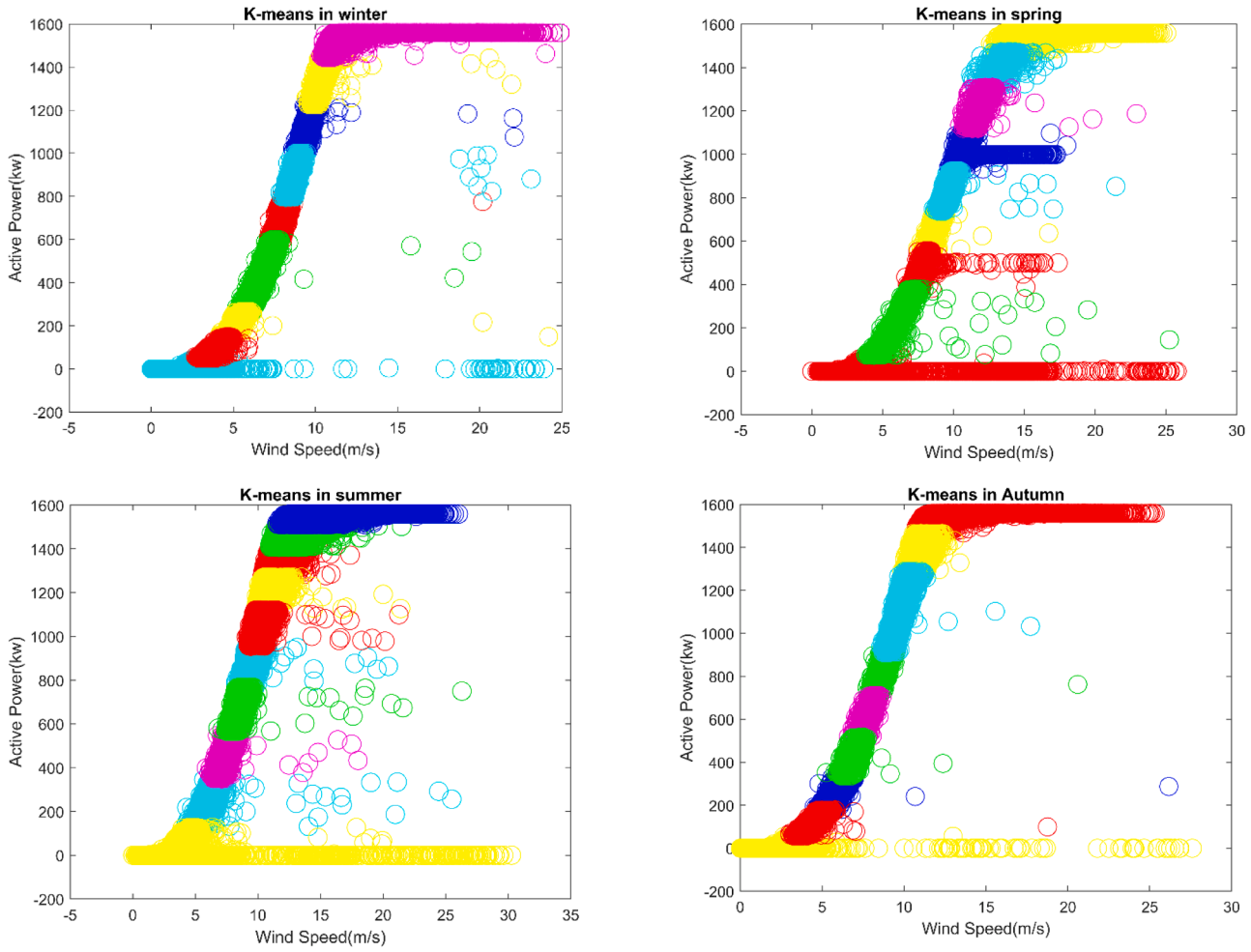

Fig. 13. Outliers detected by the K-means model by dealing with data for each season separately.

- Dynamics sensitivity: With static operating conditions, accuracy is around 96 %; it decreases to 92 % under moderate variability and to 88 % under rapid system dynamics, see it in Table 2, and Fig. 5(b).

These findings demonstrate that the proposed hybrid model effectively denoises abnormal wind-power data under varying noise and dynamic scenarios, with only gradual performance degradation as conditions worsen.

5.3.2. Comparison with denoising techniques

Let us now compare our proposed hybrid model with the denoising and robust detection methods proposed in the three referenced HVdc protection papers [24–26]

All three DC-grid schemes apply discrete wavelet transform (DWT) to denoise by decomposing DC-link signals into multiple scales, thresholding or energy-filtering the detail coefficients and then reconstructing clean transients for feature extraction. In contrast, the proposed hybrid model uses a two-stage machine-learning approach—first applying statistical filters to flag outliers, then training ensemble regressors to correct them—rather than pure time-frequency filtering. While the DWT-based methods excel at sub-millisecond transient isolation (achieving >98 % classification or location accuracy), the hybrid model delivers end-to-end cleaning of contextual anomalies in power-curve data, sustaining over 95 % cleaning accuracy under varied noise levels and dynamic conditions.

5.3.3. Comparison of the proposed hybrid model with an advanced hybrid model

The advanced hybrid model for comparison is Denoising Autoencoder (DAE) + DBSCAN + Long Short-Term Memory Autoencoder (LSTM AE). A more advanced hybrid model combines modern AI techniques to compare with our proposed hybrid model in this study: FCM + Mahalanobis distance + ANN, see it in Fig. 6.

DAE pre-cleans the data, reducing noise while preserving the structure. DBSCAN to identify dense regions (normal data) and noise (outliers) without needing cluster count and LSTM AE to learn the temporal behaviour of wind data and flag anomalies based on reconstruction error.

- The two models show good results with AUC (Area Under Curve) values close to 1, see it in Fig. 7(a), indicating strong discriminatory power. The advanced model (DAE + DBSCAN + LSTM AE) slightly outperforms our proposed hybrid model, especially in higher recall regions.
- The two models perform closely regarding Precision, Recall, and F1 Score, see it in Fig. 7(b), showing well-balanced detection. The advanced hybrid model (DAE + DBSCAN + LSTM AE) shows slightly better Recall, indicating it catches more true anomalies. Our proposed hybrid model (FCM + Mahalanobis + ANN) maintains Precision high, decreasing false positives. It confirms they can deliver similar results when properly tuned, with trade-offs in what kind of anomalies they detect better (Fig. 8).

In Table 9 although both the proposed FCM+Mahalanobis+ANN and

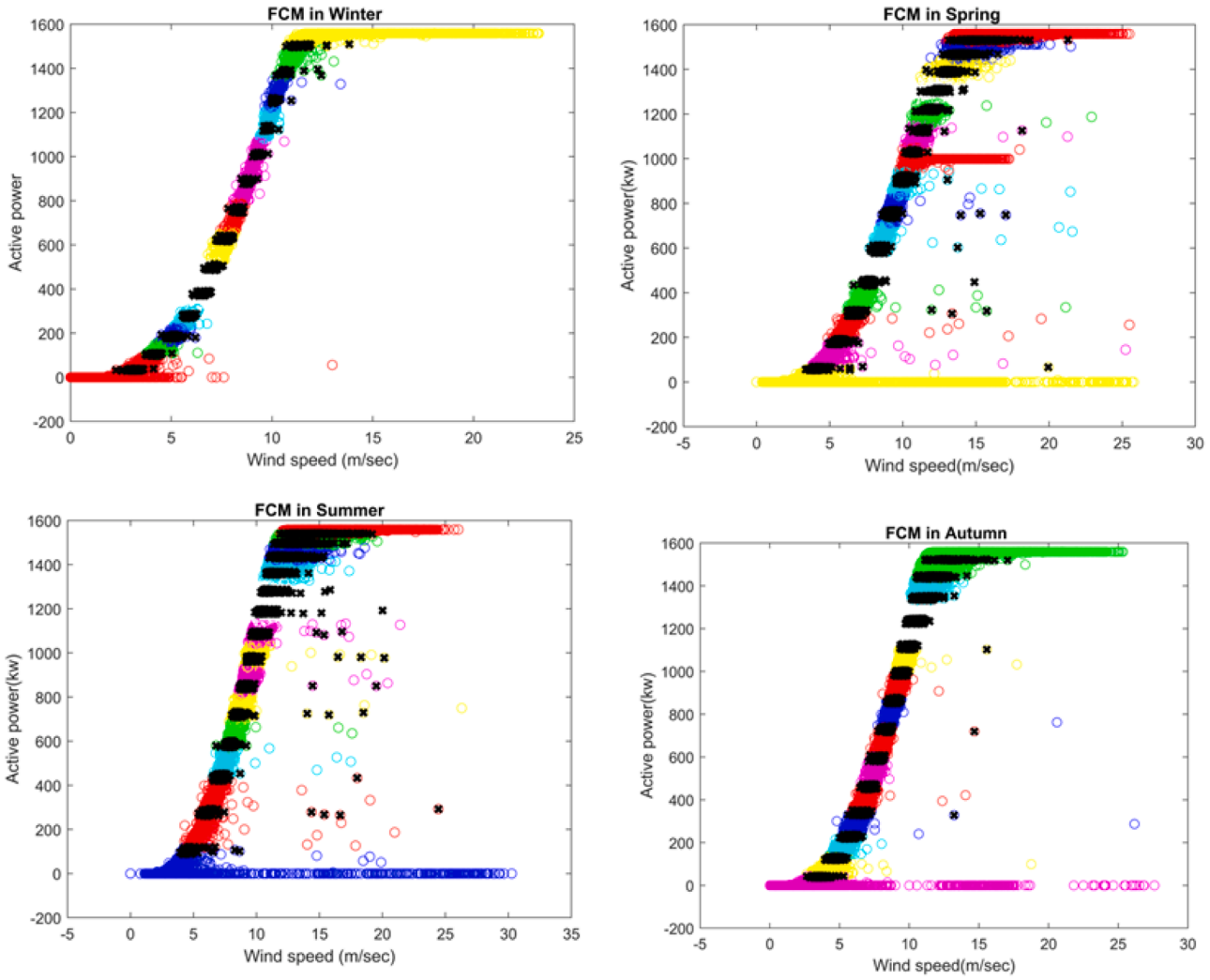

Fig. 14. Outliers detected by the FCM model by dealing with data for each season separately.

the advanced DAE+DBSCAN+LSTM AE model achieve almost flawless outlier removal and accuracy, the advanced hybrid model holds a very slight edge—RMSE drops from 0.947 to 0.9320, MAPE from 1.86 % to 1.82 %, R^2 rises from 0.980 to 0.981, and CA slips only from 1.00 to 0.99. This tiny gap reflects the autoencoder's ability to learn a richer, nonlinear latent representation of the data (and any temporal structure with the LSTM AE) before DBSCAN isolates anomalies. In contrast, the proposed model relies on covariance-scaled soft clusters in the original feature space. In practice, FCM with Mahalanobis distance already captures most multivariate outliers by down-weighting borderline points and feeding the ANN clean, homogeneous groups, so it closely approaches the deep model's performance. The autoencoder's marginal advantage lies in uncovering the most subtle, high-dimensional deviations and enforcing sequence continuity, but both models deliver virtually indistinguishable, world-class outlier detection and predictive accuracy.

Final Takeaway: The advanced hybrid model DAE + DBSCAN + LSTM AE brings a marginal improvement in reconstruction quality at the cost of one per cent in anomaly detection, whereas the simpler proposed hybrid model FCM + Mahalanobis + ANN has flawless classification with nearly identical denoising efficacy. The choice between them thus rests on whether one favours end-to-end deep representation learning or a more interpretable, clustering-based outlier rule. Also, regarding deployment needs, the FCM + Mahalanobis + ANN is lightweight and scalable, whereas the DAE + DBSCAN + LSTM AE requires more computing and maintenance.

5.4. *Visualization*

To contrast results through scatter plots, Taylor diagram, and violin plot, each plot presentation serves a distinct function based on the specific data aspect being compared. Scatter plot to compare two data sets like predictions and actuals for evaluation of correlation outliers or systematic bias.Taylor diagram evaluates the performance of various models in comparison to observations by considering standard deviation, correlation and centred RMSE.Violin Plot display data distribution or errors such as residuals of various models by merging box plot and KDE. See scatter, Taylor, and violin plots used to compare the results in Figs. 9, 10, and 11.

6. Discussion

According to the performance analysis demonstrated in Table 3 for various algorithms used to detect outliers in the wind turbine power curve across four seasons: Winter (Dec 1–Feb 28/29), Spring (Mar 1–May 31), Summer (Jun 1–Aug 31), and Autumn (Sep 1–Nov 30). The algorithm's behaviour can be highlighted in Figs. 13, 14, 15, 16, and 17.

Temporal segmentation aims to assist in comprehending and examining the fluctuations in patterns and behaviours across seasons by dealing with data for each season separately, also facilitating the

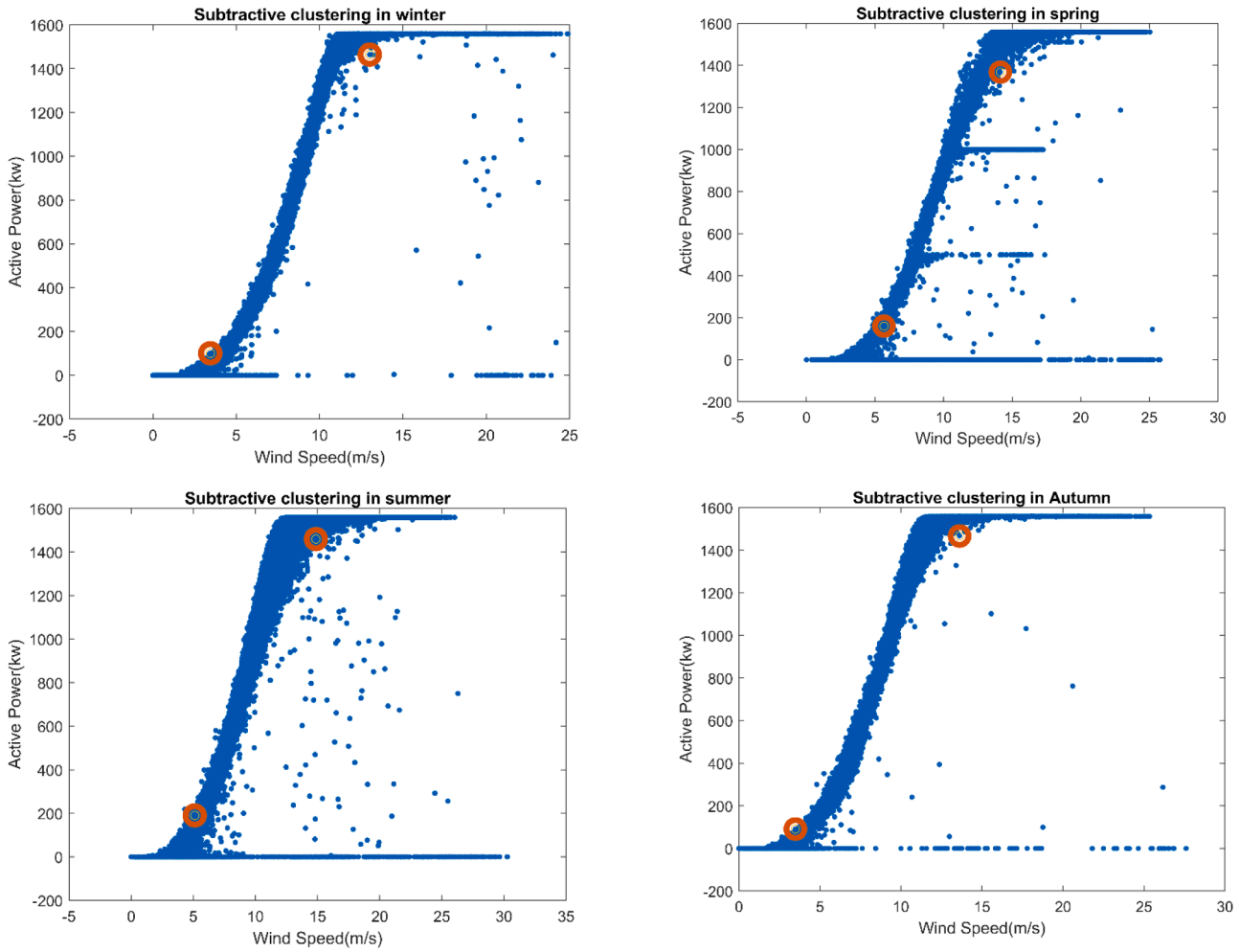

Fig. 15. Outliers detected by subtractive model through dealing with data for each season separately.

assessment of algorithms in environmental and operational settings. The variations in the seasons affect the power curve of wind turbines, given factors such as temperature variations and wind patterns alongside loads. Table 3 demonstrates that both K-means and FCM produce far lower seasonal RMSE and MAE than any of the regression- or density-based methods (SVR, ANFIS, ANN) or subtractive clustering, proving their superior ability to isolate anomalous wind-turbine readings. K-means excels at carving the data into clear, nonoverlapping groups—quickly flagging gross outliers—while FCM soft-membership assignments gracefully taper off the influence of borderline or mixed-regime points instead of forcing them into the wrong cluster. Combining these two complementary Clustering paradigms with a Mahalanobis distance filter (which weights deviations by the complete covariance structure of the inputs), the hybrid pipeline can sharply identify extreme multivariate outliers and gently down-weight subtler anomalies. Finally, feeding this cleaner, membership-weighted dataset into an ANN ensures the network learns only the authentic, noise-reduced nonlinear relationships, yielding dramatically lower prediction errors and consistently flawless outlier deletion every season. Table 4 starkly demonstrates how embedding an ANN into the Mahalanobis-based clustering framework slashes error and boosts detection accuracy: with only FCM plus Mahalanobis distance, RMSE sits at 6.33 and MAPE at 2.89 %, whereas adding the neural network drives RMSE down to 0.947 and MAPE to 1.86 %, raises R^2 from 0.96 to 0.98, and lifts combined accuracy from 0.59 to a perfect 1.00. The clustering and Mahalanobis step identifies and flags potential outliers by measuring multivariate distance. However, it cannot learn the underlying nonlinear mapping between inputs and targets—leading to large residuals when data deviate from simple cluster centroids. Introducing the ANN after clustering leverages those cleaner, outlier-filtered subgroups as structured inputs for a flexible function approximator: the network learns complex, context-specific relationships within each cluster, smooths over residual noise, and sharply down-weights or reclassifies outlier points. Therefore, this two-stage "soft-partition then nonlinear fit" workflow yields far tighter predictions and flawless outlier removal, explaining the dramatic improvements in all four key metrics. In Table 5, The Mahalanobis and Euclidean distances outperform other distance-based methods [27]. The findings in Table 5 show that the proposed hybrid model—fuzzy C-means soft clustering using Mahalanobis distance followed by an ANN—outperforms every other hybrid model not only in prediction error (RMSE 0.9475 vs 0.98–1.07 and MAPE 1.86 % vs 2.59–2.67 %) but also in its ability to identify and discard outliers (CA = 1.00 vs. 0.890–0.854). This superior outlier detection arises because Mahalanobis distance naturally flags points that lie far from the multivariate mean in any correlated feature space, and fuzzy clustering then down-weights their influence rather than forcing them into a cluster centroid—unlike Euclidean-based or hard K-means approaches, which leverage points or non-spherical data can fool. By feeding only the more homogeneous, covariance-aware clusters into the ANN, the network avoids learning spurious patterns introduced by outliers. It concentrates on authentic nonlinear relationships, yielding significantly lower residuals, tighter fit ($R^2 = 0.98$), and flawless outlier removal. In Table 6, the proposed FCM+Mahalanobis distance + ANN hybrid model not only achieves by far the best predictive accuracy

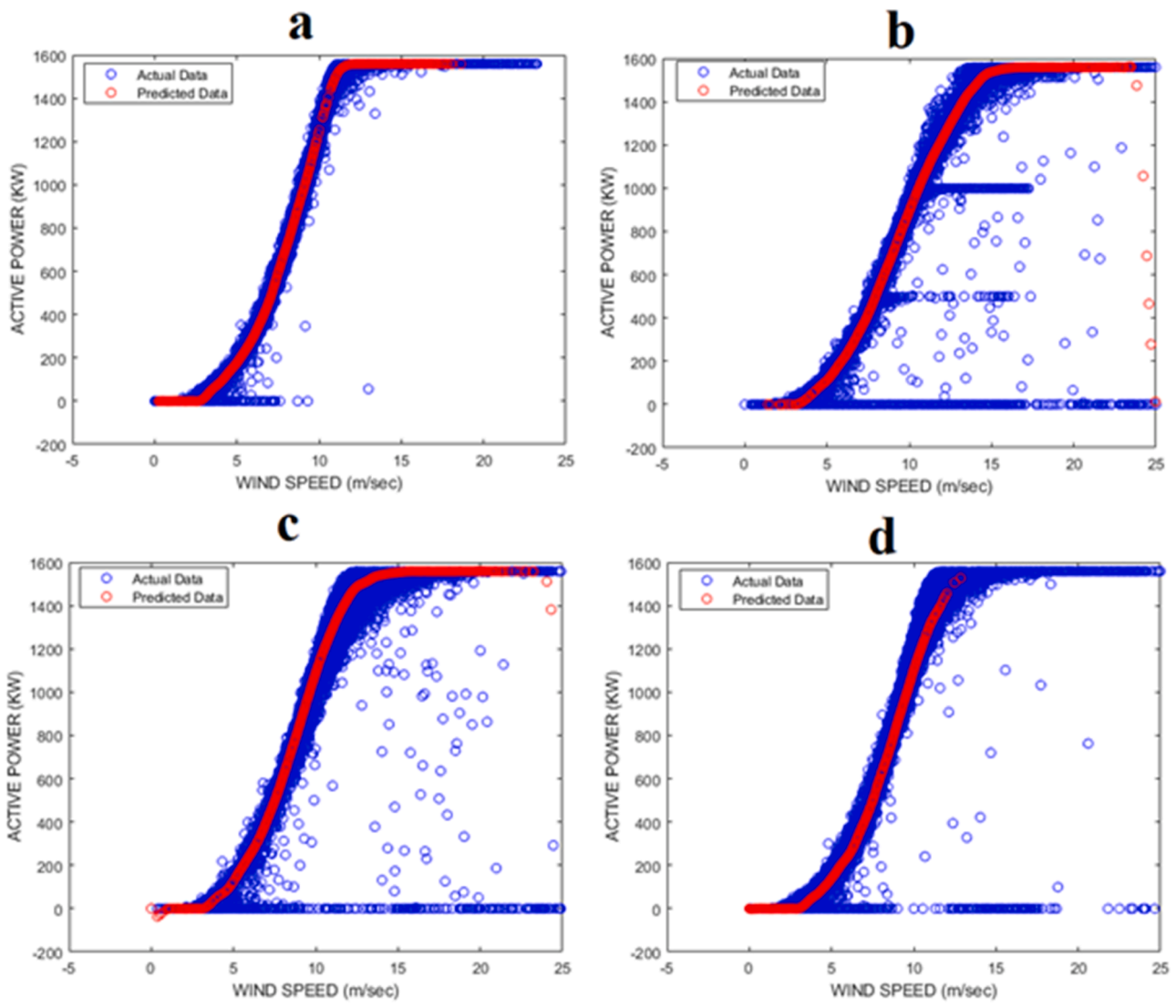

Fig. 16. Outliers detected by the SVR model by dealing with data for each season separately: a) winter, b) spring, c) summer, d) autumn.

(RMSE 0.947 vs 1.272–6.207, MAPE 1.86 % vs 2.25–8.29 %, R^2 0.98 vs 0.29–0.91) but also perfect outlier detection (CA = 1.00 vs 0.224–0.775). This superiority stems from the two-stage preprocessing: Mahalanobis distance automatically highlights points that lie anomalously far from the data's multivariate mean—capturing outliers that simple Euclidean- or density-based methods miss—and fuzzy C-means then assigns each observation a soft membership score that effectively down-weights those anomalies rather than forcing them into a single global cluster. Finally, the ANN trains on these cleaned, membership-weighted clusters,

Learning the authentic nonlinear relationships without being distorted by extreme values. In contrast, standalone methods like SVR or ANN must fit one model to all points (including outliers), and single clustering approaches (subtractive, k-means or FCM alone) either lack an explicit outlier metric or cannot balance cluster "softness" with variance-based distance, resulting in higher residual error and poorer outlier removal. Table 7 clearly shows that removing anomalous data points through the cleaning stage more than halves the prediction error and dramatically tightens the fit: RMSE falls from 2.38 to 1.27, MAPE from 11.81 % to 6.03 %, R^2 climbs from 0.83 to 0.93, and combined accuracy jumps from 0.645 to a perfect 1.00. In the pre-cleaning phase, the training set still contains sensor glitches, extreme weather spikes or other measurement artefacts that the model must attempt to learn. Hence, it over-fits to noise, yields large residuals, and cannot be generalized well across the actual operating regime. By detecting and excluding those outliers before training, the post-cleaning model focuses solely on the authentic, physically plausible patterns in turbine performance. This reduction in variance—combined with removing bias introduced by spurious readings—enables the underlying algorithm to capture the genuine nonlinear relationship between wind inputs and power output, resulting in substantially lower forecasting errors and perfect identification (CA = 1) of valid versus invalid data.

Table 8 shows that, across all four seasons, the proposed model FCM+Mahalanobis+ANN is remarkably robust to the size of the hidden layer: even with as few as 20 neurons, it achieves near-perfect R^2 (0.98–1.00) and low RMSE and MAPE, and increasing to 40–100 neurons yields only incremental gains (for example winter MAPE falls from 0.79 % at 20 neurons to 0.70 % at 80 neurons). Beyond roughly 100 neurons, the errors creep upward—MAPE edges back toward 0.90–1.00 % in winter and autumn RMSE climbs from ~0.60 to ~0.75—suggesting that large networks introduce slight overfitting without any fundamental bias reduction. In short, a moderate-sized network (around 40–100 neurons) is sufficient to capture the clusters' nonlinear structure while avoiding over-parameterization, and the model's consistently high R^2 shows that its predictive fidelity is essentially saturated over this range.

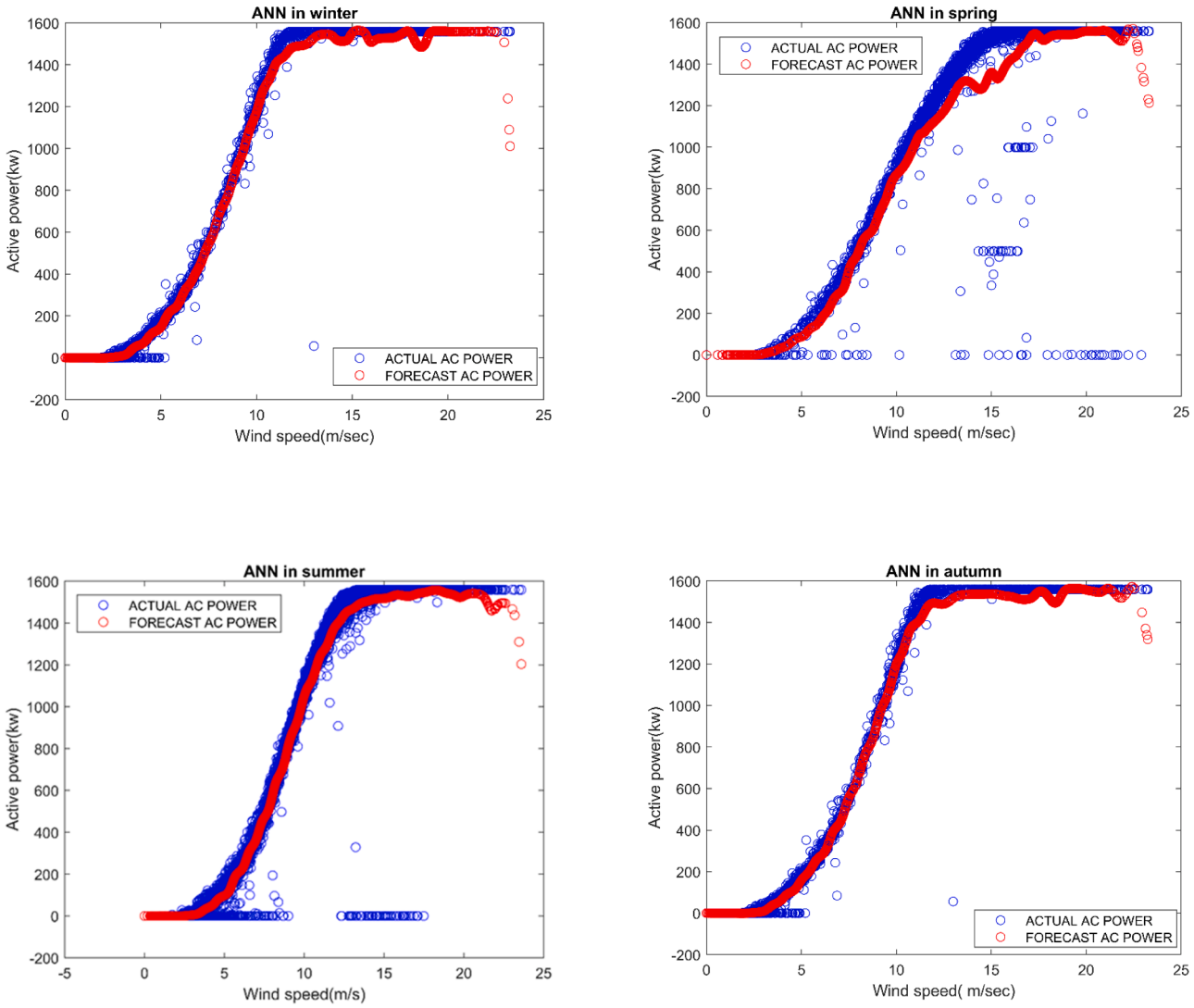

Fig. 17. Outliers detected by the ANN model by dealing with data for each season separately.

Fig. 12 illustrates that outliers are more frequent in winter, especially at higher wind speeds, which could be due to anomalies because of extreme environmental conditions or turbine behavior. There are fewer outliers in the spring, which are spread uniformly across different wind speeds. The performance is the best in summer, with no major outliers and almost a straight line. Autumn seems similar to spring but with fewer outliers at low wind speeds and more normal data distribution. The model effectively identifies and removes outliers based on FCM, Mahalanobis distance, and ANN for better accuracy in the WTPC analysis of the given data.

7. Conclusion

This paper illustrates the efficiency of incorporating Fuzzy C-Means (FCM), Mahalanobis Distance, and Artificial Neural Networks (ANN) for outlier detection and cleaning in wind data. Each of these methods is utilized to take advantage of its strengths. They are combined to leverage the ability of FCM to provide flexible clustering, the ability of the Mahalanobis Distance to detect anomalies robustly when correlations are taken into account, and the ability of the ANN to refine the final detection of outliers and identify any remaining outliers that may have been undetected by the Mahalanobis Distance. The novelty of this approach is in the ability to maintain data integrity when removing outliers systematically to produce more accurate datasets for use in other applications. The results show (1) the effectiveness of the proposed hybrid model in detecting and eliminating outliers of the wind turbine power curve and hence outperforming single methods and other traditional hybrid models; (2) improvement in the accuracy of the proposed hybrid model after the addition of ANN trained on the cleaned data; (3) while the advance model is slightly superior in reconstruction quality, our proposed hybrid model is lightweight, scalable, and remains more suitable for real-world deployment due to its low computational costs.

The novelty of this work lies in developing a two-stage outlier detection and deletion mechanism. It distinguishes this study from other studies. Most other hybrid models use a single stage to detect and remove outliers, but in this way, one cannot guarantee that all targeted outliers are detected and removed. Some unobvious outliers may be hidden in the processed data. Accordingly, our proposed hybrid model has two stages to detect and remove outliers. This process will enhance the robustness of the approach.

The pros of having cleaner wind data are important in predictive modelling and operational decision-making in wind energy and environmental sciences. More precise data sets always help optimise the accuracy of wind resource assessments, enhance the efficiency of energy forecasting models, and improve the performance of turbine operations. Hence, environmental studies dependent on wind measurements, like air quality studies and climate modelling, will also benefit from the low noise data, resulting in better analyses and policy recommendations.

This research presents a promising hybrid approach but has notable limitations. The model's multi-stage process—clustering, statistical analysis, and ANN training—introduces significant computational complexity, which can be tough with huge SCADA datasets. Additionally, the method's reliance on optimal threshold selection for Mahalanobis distance affects its generalizability, and the ANN's performance is closely tied to the quality of the initial data filtering. To improve the model's practical applicability, it is recommended that real-time deployment testing be conducted and model simplification techniques to lower computational needs explored. Cross-site validation ensures robustness across different turbine types and operating conditions.

Future work can include further developing this approach for application to other environmental data sets, e.g. solar radiation or hydrological data, where data accuracy is also important. The adaptability to various data types can be enhanced by finding other clustering algorithms, such as density-based or hybrid algorithms. Thus, the improvements in outlier cleaning methods will improve the predictive models and enhance operational efficiency in renewable energy and environmental monitoring applications.

The plan is to use the resulting cleaned data from the proposed hybrid model, FCM+Mahalanobis Distance+ANN, described in this work, in a predictive model for estimating wind turbine output power, which will be developed in subsequent work.

Subject: Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used a grammar checker to improve the grammar of this work because English is their second language. After using this grammar checker, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

CRediT authorship contribution statement

Abdelwahab Ayash Subuh: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **S. Hr. Aghay Kaboli:** Supervision. **Muhammad Waqar:** Formal analysis. **François Vallée:** Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

References

[1] Francisco Bilendo, et al., Applications and modeling techniques of wind turbine power curve for wind farms—A review, Energies 16 (1) (2023) 180, https://doi.org/10.3390/en16010180.

[2] Mingzhe Zou, Sasa Z. Djokic, A review of approaches for the detection and treatment of outliers in processing wind turbine and wind farm measurements, Energies 13 (16) (2020) 4228, https://doi.org/10.3390/en13164228.

[3] Z. Li, X.R. Luo, M.J. Liu, X. Cao, S. Du, H. Sun, Short-term prediction of the power of a new wind turbine based on IAO-LSTM, Energy Rep. 8 (2) (2022) 9025–9037, https://doi.org/10.1016/j.egyr.2022.07.030.

[4] D. Wang, M. Xu, Z. Guangming, F. Luo, J. Gao, Y. Chen, Enhancing wind power forecasting accuracy through LSTM with adaptive wind speed calibration (C-LSTM), Sci. Rep. 15 (2025), https://doi.org/10.1038/s41598-025-89398-y, 5352.

[5] M.A.A. Al-qaness, A.A. Ewees, H. Fan, L. Abualigah, A.H. Elsheikh, M.A. Elaziz, Wind power prediction using random vector functional link network with capuchin search algorithm, Ain. Shams Eng. J. 14 (5) (2022) 102095, https://doi.org/10.1016/j.asej.2022.102095.

[6] Rana Muhammad Adnan. et al. Pan evaporation estimation by relevance vector machine tuned with new metaheuristic algorithms using limited climatic data. Engineering Applications of Computational Fluid Mechanics 2023, 17(1) 2192258. https://doi.org/10.1080/19942060.2023.2192258.

[7] S. Joshi, B.L. Mahanthi, G. P, K.S. Pokkuluri, S.S. Ninawe, R. Sahu, Integrating LSTM and CNN for stock market prediction: A dynamic machine learning approach, J. Artif. Intell. Technol. V (5) (2025), https://doi.org/10.37965/jait.2025.0652.

[8] Z. Wang, W. Liu, X. Wang, Abnormal data cleaning of wind turbine power curve using bayesian change point-quartile combined algorithm, Proc. Inst. Mech. Eng. A: J. Power Energy 237 (5) (2023) 495–503, https://doi.org/10.1177/09576509221119563.

[9] F. Zhang, X. Zhang, Z. Xu, K. Dong, Z. Li, Y. Liu, Cleaning of abnormal wind speed power data based on quartile RANSAC regression, Energies 17 (22) (2024) 5697, https://doi.org/10.3390/en17225697.

[10] J.Y. Park, J.K. Lee, K.Y. Oh, J.S Lee, Development of a novel power curve monitoring method for wind turbines and its field tests, IEEE Trans. Energy Convers. 29 (1) (2014) 119–128, https://doi.org/10.1109/TEC.2013.2294893.

[11] C. Monteiro, et al., Wind Power Forecasting: State-of-the-Art 2009 (No. ANL/DIS-10-1), Argonne National Laboratory (ANL), United States, 2009, https://doi.org/10.2172/968212.

[12] A. Kusiak, H. Zheng, Z. Song, On-line monitoring of power curves, Renew. Energy 34 (6) (2009) 1487–1493, https://doi.org/10.1016/j.renene.2008.10.022.

[13] X. Shen, X. Fu, C. Zhou, A combined algorithm for cleaning abnormal data of wind turbine power curve based on change point grouping algorithm and quartile algorithm, IEEe Trans. Sustain. Energy 10 (1) (2019) 46–54, https://doi.org/10.1109/TSTE.2018.2822682.

[14] E. Taslimi-Renani, M. Modiri-Delshad, M.F.M. Elias, N.A. Rahim, Development of an enhanced parametric model for wind turbine power curve, Appl. Energy 177 (2016) 544–552, https://doi.org/10.1016/j.apenergy.2016.05.124.

[15] L. Morissette, S. Chartier, The k-means clustering technique: general considerations and implementation in mathematical, Tutor. Quant. Methods Psychol. 9 (1) (2013) 15–24, https://doi.org/10.20982/tqmp.09.1.p015.

[16] Bezdek, J.C, "Pattern recognition with fuzzy objective function algorithms. New York, Springer Science & Business Media, 1981. DOI:10.1007/978-1-4757-0450-1.

[17] M. Zou, S.Z. Djokic, A review of approaches for the detection and treatment of outliers in processing wind turbine and wind farm measurements, Energies 13 (16) (2020) 4228, https://doi.org/10.3390/en13164228.

[18] R.K. Pandit, D. Infield, Comparative assessments of binned and support vector regression-based blade pitch curve of a wind turbine for the purpose of condition monitoring, Int. J. Energy Env. Eng. 10 (2019) 181–188, https://doi.org/10.1007/s40095-018-0287-3.

[19] K. Levenberg, A method for the solution of certain non-linear problems in least squares, Q. Appl. Math. 2 (2) (1944) 164–168, https://doi.org/10.1090/qam/10666.

[20] D.W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters, Soc. Ind. Appl. Math. 11 (2) (1963) 431–441. https://www.jstor.org/stable/2098941.

[21] Shamshirband Sh, et al., Adaptive neuro-fuzzy optimization of wind farm project net profit, Energy Convers. Manage 80 (2014) 229–237, https://doi.org/10.1016/j.enconman.2014.01.038.

[22] Petković D, Cojbašić Ž, Nikolić V. Adaptive neuro-fuzzy approach for wind turbine power coefficient estimation. Renew and sustain energy rev 2013, 28, 191–195. https://doi.org/10.1016/j.rser.2013.07.049.

[23] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, Int. J. Forecast. 22 (4) (2006) 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

[24] M.Z. Yousaf, S. Khalid, M.F. Tahir, A. Tzes, A. Raza, A novel dc fault protection scheme based on intelligent network for meshed dc grids, Int, J. Electr, Power Energy Syst 154 (2023), https://doi.org/10.1016/j.ijepes.2023.109423. Article 109423.

[25] M.Z. Yousaf, H. Liu, A. Raza, A. Mustafa, Deep learningbased robust dc fault protection scheme for meshed HVdc grids, CSEE J. Power Energy Syst 9 (6) (2023) 2423–2434, https://doi.org/10.17775/CSEEJPES.2021.03550.

[26] M.Z. Yousaf, M.F. Tahir, A. Raza, M.A. Khan, F. Badshah, Intelligent sensors for dc fault location scheme based on optimized Intelligent architecture for HVdc systems, Sensors 22 (24) (2022) 9936, https://doi.org/10.3390/s22249936.

[27] Raúl Ruiz de la Hermosa González-Carratoa, "Wind farm monitoring using Mahalanobis distance and fuzzy clustering" renewable energy 2018, 123,526-540. https://doi.org/10.1016/j.renene.2018.02.097.

[28] Shenglei Pei, Yifen Li, Wind turbine power curve modeling with a hybrid machine learning technique, Appl. Sci. 9 (22) (2019) 4930, https://doi.org/10.3390/app9224930.

Abdelwahab Ayash Subuh is currently a PhD student in the Power Systems and Markets Research Group at the University of Mons, Belgium, specializing in renewable energy. His research focuses on data-driven methods for wind turbine power output forecasting and monitoring. He holds a master's degree in electrical engineering and a bachelor's degree in electrical engineering.

Before pursuing academia, he worked as a control engineer at Baiji Gas Power Plant, Iraq (2005– 2014), where he gained extensive experience with TXP control systems for SIEMENS V94.2 gas turbines. He has also undergone specialized training in pulse filter systems operation and maintenance by Rotring company in Germany and TXP control systems by SIEMENS experts at Baiji power plants.

His computer skills include Python, MATLAB (MathWorks Inc.), and C++.