Amirkabir Journal of Mechanical Engineering

Amirkabir J. Mech. Eng., 57(10) (2026) 1263-1286
DOI: 10.22060/mej.2026.24688.7894

A Data-Driven Framework for Wind Turbine Power Curve Cleaning and Abnormal Data detection Based on Binning and Quantiles

Alireza Aghajani Mobarakeh, Javad Poshtan*

Control Group, Department of Electrical Engineering, Iran University of Science and Technology, Tehran, Iran.

ABSTRACT: In this study, a preprocessing approach is proposed to improve training accuracy and computational efficiency by leveraging SCADA data and integrating machine learning algorithms with statistical techniques for unlabeled data. The method reduces the training dataset substantially by partitioning the data into equal numbers of data points and selecting representative samples based on quantiles. Using this strategy, the model requires only 0.2% of the original dataset for training in this study. Subsequently, abnormal data points are identified using a power curve model using quantile thresholds. The proposed method is evaluated against DBSCAN and a KNN-based model. Experimental results obtained from real-world wind farm data indicate that the proposed method combined with KNN outperforms the DBSCAN method. Specifically, both MAE and RMSE decreased by approximately 15%, reflecting improved predictive accuracy. Computationally, the execution time of the proposed approach was about 0.15 seconds, compared to 0.99 seconds for the DBSCAN method, corresponding to a runtime reduction exceeding 50%. Moreover, unlike DBSCAN, which requires precise parameter tuning or additional constraints tailored to the power curve structure when dealing with dense or linear outliers, the proposed approach is capable of automatically eliminating outlier data points from the wind turbine power curve without the need for predefined filters or explicit boundary definitions prior to the cleaning process.

Review History:

Received: Sep. 05, 2025
Revised: Dec. 14, 2025
Accepted: Apr. 16, 2026
Available Online: Apr. 30, 2026

Keywords:

Data Cleaning
Wind Turbine Power Curve (WTPC)
Machine Learning (ML)
Supervisory Control and Data Acquisition (SCADA)

1- Introduction

Condition monitoring and abnormal data identification function as a fundamental and essential protective layer, preventing undesirable events, ensuring human safety, and safeguarding the environment.

With the significant growth of renewable energy in recent years, condition monitoring has become increasingly important in energy sector. In particular, wind energy has gained significant attention as a source of renewable energy, due to its availability and low environmental impact. Therefore, condition monitoring and timely detection of abnormal data as potentially faulty data, along with early alarms to operators, can help reduce costs and enhance system reliability [2 ,1] . The power curve of wind turbines can serve as a pivotal key indicator for detecting abnormalities. For instance, data exhibiting a behavior that differs significantly from the power curve data can be identified as abnormal data.

In most studies on wind turbine condition monitoring based on power curves, it is common to apply filters before the preprocessing step in order to prepare high-quality and clean data for model training. Recent approaches combine statistics and machine learning (ML) to improve accuracy, but they usually involve multiple complex preprocessing steps [3, 4]. Therefore, this paper introduces a computationally efficient and cost-effective method based on machine learning and quantiles that identifies and removes abnormal data without the need for manual filters during the preprocessing step, with low dependence on the turbine's structural characteristics.

2- Methodology

The wind turbine power curve, which represents the non-linear relationship between wind speed and output power, is a key indicator in analyzing the efficiency of wind energy systems. Equation (1) represents this relationship, where p is the output power of the turbine.

$$p = \frac{1}{2}\rho A C_p(\lambda,\beta)v^3 \quad (1)$$

Various methods have been proposed for modeling Equation (1), such as parametric and non-parametric approaches. Among these approaches, machine learning methods are superior due to their capacity to accurately

*Corresponding author's email: jposhtan@iust.ac.ir

Copyrights for this article are retained by the author(s) with publishing rights granted to Amirkabir University Press. The content of this article is subject to the terms and conditions of the Creative Commons Attribution 4.0 International (CC-BY-NC 4.0) License. For more information, please visit https://www.creativecommons.org/licenses/by-nc/4.0/legalcode.

approximate complex realworld systems with high accuracy and operate as blackbox models. In this study, the K-Nearest Neighbors (KNN) algorithm is implemented to learn the relationship. It estimates the output using the k nearest observations within the feature space.

The first step in preparing the data for the model training involves cleaning the dataset by removing outliers. A popular method is Density-Based Spatial Clustering of Applications with Noise (DBSCAN). Equation (2) describes the main logic behind the DBSCAN algorithm.

$$N_{\varepsilon}(p) = \{q \in \mathrm{D} \mid \mathrm{dist}(p,q) \le \varepsilon\} \tag{2}$$

In Equation (2), N represents the set of neighbors of the point p, consisting of all q points with a distance smaller than ε. If the size of set N exceeds a specified threshold (Minpts), point p is considered a normal point. Otherwise, p is an outlier (Equation (3)).

$$|N_{\varepsilon}(p)| \ge \mathrm{Minpts} \tag{3}$$

Although DBSCAN can efficiently detect outliers, it still requires additional predefined filters to remove some initial data.

Subsequently, a KNN model is trained on the preprocessed data, which uses a threshold to find any significant deviation from the outliers-free power curve (Equation *(5)* ,*(4)*). In these Equations, r is the residual signal, x_t represents the data point at time t, and σ is the standard deviation. If the residual values exceed the acceptable bounds of the power curve, they can be flagged as potential abnormal data.

$$r_t = x_t - \mathrm{Predicted\ value} \tag{4}$$

$$|r_t| \ge 3 \times \sigma_r \tag{5}$$

One of the limitations of Equation *(5)* is the assumption that the residual signal is normally distributed. The real distribution of a wind turbine residual signal is not exactly normal, and using a 3σ threshold can lead to false alarms, as the threshold ignores other factors, such as skewness.

The proposed data-driven approach uses the actual residual signal distribution to determine the appropriate threshold and "avoids common simplifications, such as assuming normality.

The proposed method has two main steps:

Selecting center points and cleaning data using quantile thresholds.

Defining threshold bounds for detecting abnormal data.

The initial stage involves data partitioning according to IEC standards and the selection of center points. Overall, after selecting wind speed as the appropriate feature using the correlation matrix, centers can be selected. A key indicator for the selection is mean value within each bin. While the mean value is heavily sensitive to outliers, quantiles might provide a better solution, as they are more robust and allow focusing on specific parts of the data distribution, particularly the higher values on the power curve that indicate stable and optimal turbine performance. This provides a more accurate representation of the turbine's maximum power and actual operating conditions. Therefore, the wind speed data is divided into bins with approximately equal sample counts in each interval, in order to examine the power curve behavior uniformly. Subsequently, the median points are chosen as selected data points.

Using the preprocessed data, a KNN model is trained. Then, the real-time data is compared with the model prediction, and if the difference is significant, the data is flagged as abnormal data and potentially faulty data. Equation *(6)* shows the threshold relationship where r is the residual and Qs are random thresholds for the left and right side of the residual signal which can be selected based on the SCADA data structure.

$$|r_{\mathrm{Rt}}| \succ \mathrm{Q}_{\mathrm{Rp,}} \quad |r_{\mathrm{Lt}}| \succ \mathrm{Q}_{\mathrm{Lp,}} \tag{6}$$

In summary, the data is cleaned and the residual signal is stored. This cleaning is performed using binning and quantile thresholding. Then, the ML model is trained on this cleaned data. Given the residual signal and the local quantile threshold (for each interval), abnormal conditions are identified.

3- Results

The experimental results are based on a dataset collected from a real wind turbine located in Turkey. In the first step, outliers in the raw dataset are removed with both DBSCAN and the proposed algorithms to enable performance comparison.

Although the full dataset is available, only 100 centers of the power curve are selected using the proposed method to train the model. After training, the proposed method demonstrates higher accuracy and shorter execution time than DBSCAN method (Fig. 1).

Finally, local quantile-based thresholds are applied to detect the abnormalities (Fig. 2).

4- Conclusion

This study introduces a framework for cleaning wind turbine power curve data and detecting abnormalities. The proposed preprocessing step involves segmenting wind speed into bins and selecting the median of each bin as a center point to train the model, which effectively reduces computational load while preserving the accuracy of power curve modeling. For abnormal data detection, the method employs local quantile thresholds for each bin instead of a global 3σ threshold, thereby removing the assumption of normality. Experimental results demonstrate that this framework efficiently removes outliers, reduces execution time, and provides a more accurate power curve model compared to

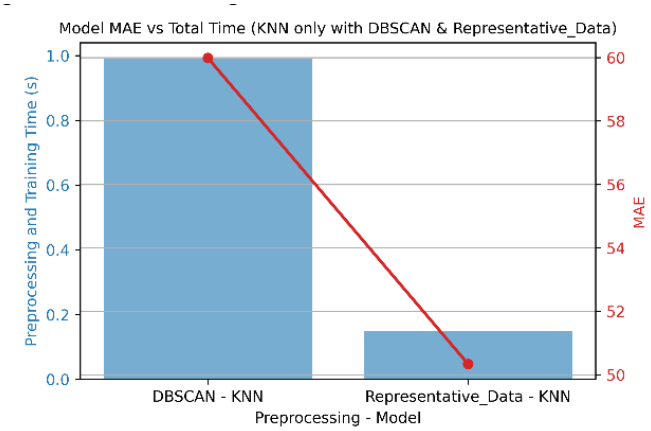


Fig. 1. Time and MAE decrease chart.

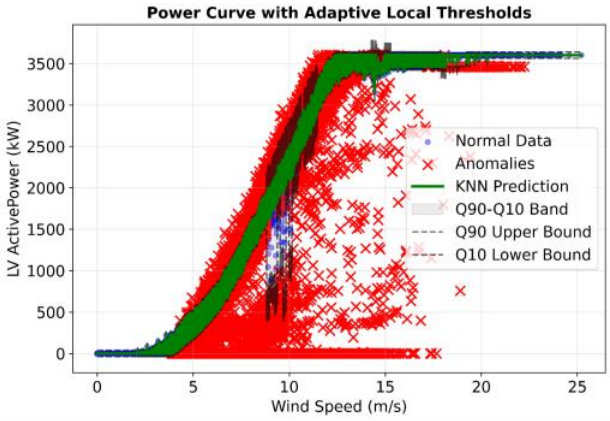


Fig. 2. Power curve and abnormal data.

the DBSCAN method, without requiring additional filters. Future research directions include integrating this framework with more advanced algorithms, utilizing environmental data, and employing multi-input power curves.

References

[1] C. Global Wind Energy, Global wind report 2023, Lisbon, 2023.

[2] C. Global Wind Energy, Global wind report 2024, Lisbon, 2024.

[3] C. Zhang, D. Hu, T. Yang, Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost, Reliability Engineering and System Safety, 222 (2022) 108445–108445.

[4] L. Xiang, X. Yang, A. Hu, H. Su, P. Wang, Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks, Applied Energy, 305 (2022) 117925–117925.

نشریه مهندسی مکانیک امیرکبیر

نشریه مهندسی مکانیک امیرکبیر، دوره ۵۷، شماره ۱۰، سال ۱۴۰۴، صفحات ۱۲۶۳ تا ۱۲۸۶
DOI: 10.22060/mej.2026.24688.7894

چارچوبی داده‌محور برای پاک‌سازی منحنی توان توربین بادی و شناسایی داده‌های نابهنجار بر اساس بازه‌بندی و چندک

علیرضا آقاجانی مبارکه، جواد پشتان *

گروه کنترل، دانشکده مهندسی برق، دانشگاه علم و صنعت ایران، تهران، ایران.

خلاصه: در این پژوهش، به‌منظور افزایش دقت و سرعت آموزش، باتکیه‌بر داده‌های اسکادا و ترکیب الگوریتم‌های یادگیری ماشین با روش‌های آماری، رویکردی برای پیش‌پردازش داده‌های بدون برچسب و حذف خودکار داده‌های پرت ارائه شده است. در این روش، با تقسیم‌بندی داده‌ها به بازه‌های مساوی و انتخاب داده‌ی نماینده بر اساس چندک، حجم داده‌های آموزشی به طور چشمگیری کاهش می‌یابد (در این پژوهش تنها با ۰/۲٪ کل داده‌ها). سپس با استفاده از مدل منحنی توان و حدود آستانه‌ی چندکی محلی، داده‌های نابهنجار شناسایی می‌گردند. روش پیشنهادی با الگوریتم دی‌بی‌اسکن و K نزدیک‌ترین همسایه مقایسه شده است. نتایج تجربی روی داده‌های واقعی مزرعه بادی نشان می‌دهد که روش پیشنهادی در ترکیب با K نزدیک‌ترین همسایه عملکرد بهتری نسبت به دی‌بی‌اسکن دارد؛ به طور خاص، هر دو مقادیر میانگین خطای مطلق و خطای جذر میانگین مربعات حدود ۱۵٪ کاهش‌یافته و بیانگر خطای پیش‌بینی پایین‌تر است. از نظر محاسباتی نیز، زمان اجرای روش پیشنهادی برابر ۰/۱۵ ثانیه و در دی‌بی‌اسکن ۰/۹۹ ثانیه گزارش شده است و زمان اجرا با روش معرفی شده، بیش از ۵۰٪ کاهش به همراه داشته است. افزون بر این، برخلاف دی‌بی‌اسکن که در برابر داده‌های پرت متراکم یا خطوار نیازمند تنظیمات دقیق یا تعریف دقیق شرایط اضافه بر اساس ساختار منحنی توان است روش پیشنهادی بدون نیاز به تعریف فیلتر و مرزبندی‌های خاص قبل از اعمال روش پاک‌سازی، قادر به حذف خودکار داده‌های غیرواقعی و پرت در منحنی توان توربین بادی می‌باشد.

تاریخچه داوری:
دریافت: ۱۴۰۴/۰۶/۱۴
بازنگری: ۱۴۰۴/۰۹/۲۳
پذیرش: ۱۴۰۴/۰۱/۲۷
ارائه آنلاین: ۱۴۰۵/۰۲/۱۰

کلمات کلیدی:
پاک‌سازی
منحنی توان توربین بادی
یادگیری ماشین
سامانه نظارتی و گردآوری داده
تشخیص داده‌ی نابهنجار

۱- مقدمه

شناسایی داده‌های نابهنجار و عیب همانند یک‌لایه حفاظتی ابتدایی و مهم عمل می‌کند که با جلوگیری از وقوع رخدادهای ناخواسته، تضمین‌کننده سلامت انسان‌ها و صیانت از محیط است. تشخیص عیب می‌تواند قبل از تبدیل اختلالات کوچک به بحران‌های بزرگ و پرهزینه، آن‌ها را آشکار سازد و از حوادث جدی و حتی جزئی، جلوگیری کند. در کنار عیب، داده‌های نابهنجار وجود دارند که ممکن است عیوب بالقوه‌ای باشند و به همین علت، اهمیت شناسایی آن‌ها پررنگ می‌شود. نکته‌ی محدودکننده دررابطه‌با این نوع داده‌ها، عدم وجود برچسب برای صحت سنجی وقوع عیب است؛ بنابراین این داده‌ها ممکن است یک عیب جدی باشند یا برعکس، نشان‌دهنده‌ی وقوع شرایط سالم و نرمال در توربین بادی باشند. کشف به‌موقع داده‌های نابهنجاری به‌عنوان عیوب بالقوه در سامانه و هشدار اولیه به متخصص بهره‌برداری می‌تواند به کاهش هزینه و افزایش قابلیت اطمینان سامانه کمک می‌کند. فرایند تشخیص این‌گونه داده‌ها موجب کاهش مخارج نگهداری و بهره‌برداری می‌شود و به حرکت به سمت نگهداری پیشگیرانه (اقدام مؤثر و به‌موقع قبل از وقوع عیب) سرعت می‌بخشد. همچنین، بهره‌گیری از حسگرهای هوشمند و سامانه‌های پردازشی پیشرفته، امکان تحلیل دقیق حجم گسترده‌ی داده‌ها را فراهم آورده است. همگام با پیشرفت این فناوری‌ها به‌ویژه توسعه روش‌های یادگیری ماشین و پایش توربین، تشخیص داده‌های نابهنجار و عیب دیگر محدود به بازرسی‌های میدانی و دستی نبوده و تشخیص داده‌ی هوشمند به یکی از ارکان اصلی ایمنی و بهره‌وری در صنایع بدل شده است.

ازسوی‌دیگر، نقش روبه‌رشد توربین بادی در صنعت انرژی جهانی به‌عنوان یکی از پایه‌های تولید برق تجدیدپذیر، اهمیت شناسایی داده‌ی نابهنجار در این تجهیزات را دوچندان می‌سازد. انرژی‌های تجدیدپذیر شامل انرژی بادی، به دلیل دردسترس‌بودن و آلایندگی اندک، جایگاه

* نویسنده عهده‌دار مکاتبات: jposhtan@iust.ac.ir

حقوق مؤلفین به نویسندگان و حقوق ناشر به انتشارات دانشگاه امیرکبیر داده شده است. این مقاله تحت لیسانس آفرینندگی مردمی (Creative Commons License) در دسترس شما قرار گرفته است. برای جزئیات این لیسانس، از آدرس https://www.creativecommons.org/licenses/by-nc/4.0/legalcode دیدن فرمائید.

مهمی در رویکردهای کلان دولت‌ها داشته است. در سال‌های اخیر، حجم سرمایه‌گذاری در این بخش افزایش قابل‌توجهی یافته است و همچنان روبه‌افزایش پیش‌بینی می‌شود [۱, ۲]. از بین روش‌های تولید انرژی پاک نظیر خورشید، بادی و...، انرژی بادی به دلیل تأثیر مخرب کم‌تر بر محیط و عملکرد بهتر، جایگاه برجسته‌تری یافته است. سهم انرژی بادی از بازار انرژی جهانی به‌صورت صعودی و در حدود ۱۰ درصد است [۱]. از همین جهت، پایش وضعیت توربین بادی به محور اصلی بسیاری از پژوهش‌های اخیر تبدیل شده است [۳].

اولین گام در پایش وضعیت هوشمند توربین بادی مدل‌سازی است؛ ولی قبل از به‌دست‌آوردن مدل، جهت تخمین رفتار سامانه، لازم است دسترسی به داده‌های مناسب آموزش وجود داشته باشد؛ بنابراین اگر دقیق‌تر نگاه شود اولین گام اصلی در پایش وضعیت توربین بادی، پاک‌سازی داده و آماده‌سازی داده‌ی مناسب است و پیش‌پردازش داده‌ها نقش حیاتی در آموزش مدل‌های پیش‌بینی و دقت مدل ایفا می‌کنند به صورتی که داده‌های کم‌کیفیت قطعاً به‌دقت مدل لطمه‌ی جدی وارد می‌کند.

روش‌های متنوعی جهت پاک‌سازی داده‌های پرت و افزایش کیفیت داده‌ها در توربین بادی وجود دارد. روش‌های سنتی پاک‌سازی داده در توربین بادی عمدتاً بر قواعد آماری ساده و مدل‌های فیزیکی مبتنی هستند. برای مثال، یکی از پرکاربردترین این روش‌ها، روش بازه‌بندی است که داده‌ها را بر اساس سرعت باد دسته‌بندی و میانگین توان تولیدی هر بازه محاسبه می‌شود [۴] و از این طریق به پاک‌سازی و مدل‌سازی می‌پردازد. این رویکرد به‌عنوان «روشی قابل‌اعتماد و تکرارپذیر برای محاسبه دقیق منحنی توان توربین بادی» شناخته شده است [۵] و بعضاً از بازه‌های ۰٫۵ متر بر ثانیه استفاده می‌شود که منطبق بر استاندارد معرفی شده توسط IEC است [۴].

باوجود عملکرد نسبتاً مناسب روش‌های آماری، این روش‌ها در برابر داده‌های پرت و الگوهای پیچیده مقاوم نیستند و در مدل‌سازی روابط غیرخطی بین سرعت باد و توان محدودیت دارند [۶]. همچنین، روش‌های آماری سنتی قادر به مدل‌سازی طبیعت بسیار تصادفی شرایط باد و حالات عملیاتی پیچیده توربین‌ها نیستند [۷].

یکی از معروف‌ترین روش‌های آماری مدرن جهت پاک‌سازی داده، الگوریتم بدون نظارت خوشه‌بندی فضایی مبتنی بر چگالی یا به‌اختصار دی‌بی‌اسکن است که به دلیل شناسایی نقاط پرت، بدون نیاز به تعداد دقیق خوشه‌ها موردتوجه قرار گرفته است [۸, ۹]. نکته‌ی قابل‌تأمل درباره‌ی دی‌بی‌اسکن، ناتوانی آن هنگام مواجه با داده‌های ناهمگون یا با چگالی متفاوت است [۹]. این ضعف، به‌ویژه خود را در داده‌های اسکادا در توربین‌های بادی با حالات عملیاتی متنوع و شرایط محیطی مختلف نشان می‌دهد و برای الگوریتم دی‌بی‌اسکن چالش‌آفرین است [۱۰]. برای رفع این‌گونه محدودیت‌ها، رویکردهای ترکیبی مانند ترکیب جنگل ایزوله و دی‌بی‌اسکن توسعه یافته‌اند و در حذف داده‌های پرت و بازسازی منحنی توان نتایج بهتری ارائه می‌کنند [۱۱]؛ بنابراین اخیراً، توجه به الگوریتم‌های مناسب‌تر جهت پوشش ضعف‌های دی‌بی‌اسکن معطوف شده است. در سال‌های اخیر همچنین تمرکز ویژه‌ای بر روش‌های آماری مقاوم از جمله رگرسیون چندکی شده است. این روش قادر است چندین صدک از توزیع توان را هم‌زمان مدل کند و تصویر کامل‌تری از رفتار نرمال توربین ارائه دهد [۱۲, ۱۳]. استفاده از چندک‌ها نه‌تنها در شناسایی نابهنجاری مؤثر است؛ بلکه در انتخاب داده‌های مناسب برای آموزش مدل و کاهش هزینه‌های محاسباتی نیز کاربرد دارد [۱۴, ۱۵]. همچنین، تکنیک‌های پردازش تصویر برای پاک‌سازی داده‌ها به کار گرفته شده‌اند؛ در این روش‌ها، منحنی‌های توان به‌صورت تصویر در نظر گرفته می‌شوند و الگوریتم‌های طبقه‌بندی برای شناسایی نابهنجاری‌ها اعمال می‌شوند [۱۶, ۱۷]. همچنین برخی دیگر از حدود آستانه‌ی تصویری یا حذف نقاط پرت به‌وسیله‌ی تحلیل پیکسل‌ها اقدام کرده‌اند [۱۷, ۱۸]. برخی از تحقیقات معاصر بر توسعه چارچوب‌های خودکار برای پردازش بلادرنگ داده‌های اسکادا بادقت بالا تمرکز دارند. ادغام معماری‌های یادگیری عمیق مانند شبکه‌های LSTM همراه با رمزگذارهای خودکار حذف نویز، قابلیت شناسایی وابستگی‌های زمانی و الگوهای پیچیده را فراهم می‌کند [۱۹]. این سامانه‌ها قادرند علاوه بر پاک‌سازی داده‌های تاریخی، پایش بلادرنگ وضعیت توربین‌ها را نیز ارائه دهند [۲۰] و شامل آستانه‌های تطبیقی و مکانیزم‌های خود یادگیر هستند [۲۱]. برخی پژوهشگران با بهره‌گیری از مدل رفتار نرمال چندهدفه، الگوهای مرجع برای متغیرهای اصلی اسکادا ایجاد کردهاند تا انحرافات کل سامانه شناسایی شود [۲۲]. مرجع [۲۳] با ترکیب شبکه عصبی پس انتشار و ارزیابی فازی، به پیش‌بینی خطاهای عمومی در گیربکس پرداخته است. برای یاتاقان اصلی، روش فاصله ماهالانوبیس در [۲۴] و الگوریتم پیش‌آگهی مبتنی بر استخراج ویژگی در [۲۲]، امکان تشخیص زودهنگام و پیش‌بینی زمان خرابی را فراهم کرد. مدل‌سازی منحنی توان به‌عنوان شاخص سلامت سامانه نیز با توابع پایه شعاعی در [۲۳] و برای یخ‌زدگی پره‌ها با مدل تشخیصی یخ در [۲۴]، بدون نیاز به برچسب‌گذاری، تنها بر پایه توزیع داده‌های عملیاتی صورت‌گرفته است.

در اکثر قابل‌توجهی از کارهای موجود در پایش وضعیت توربین بادی با داده‌های بدون برچسب و به‌خصوص بر مبنای منحنی توان، چه روشهای پارامتریک و غیرپارامتریک، بخشی از پیش‌پردازش به‌صورت دستی انجام می‌شود؛ یعنی برای دی‌بی‌اسکن به‌صورت دستی فیلتری اضافه می‌شود تا هنگام مقابله با داده‌های پرت متراکم خطشکل در توربین بادی، عملکرد مناسبی ارائه دهد. به‌عبارت‌دیگر، برای بازههای معینی از سرعت باد توربین بادی، دادههای داده‌ی این نوع داده‌های پرت با استفاده از فیلتر حذف میشود و سپس پاک‌سازی نهایی به الگوریتمهایی نظیر دی‌بی‌اسکن محول میشود. توسعه‌های اخیر بر ادغام چندین روش برای ایجاد چارچوب‌های پایدار پاک‌سازی داده تمرکز دارند. ترکیب یادگیری ماشین با آمار، راهکاری نویدبخش برای غلبه بر محدودیت روش‌های سنتی و همزمان افزایش دقت پایش وضعیت توربین است [۱۹, ۲۵] ولی این روش‌های ترکیبی، اغلب شامل مراحل متعدد پیش‌پردازش داده، از جمله شناسایی اولیه نقاط پرت، خوشه‌بندی و اعتبارسنجی مبتنی بر چندک‌ها هستند. نیاز به یک الگوریتم دقیق و سریع که بهصورت وابستگی شدید به ساختار توربین بادی و بدون نیاز به فیلتر، داده‌های پرت را حذف نماید و همزمان ساده و ارزان باشد باعث شد که روش جدیدی در این مقاله معرفی گردد.

برای حل چالش‌های مطرح شده، کلیت و خلاصه‌ی مقاله در دو بخش کلی، به شرح زیر خواهد بود:

۱- هدف به‌دست‌آوردن یک منحنی توان فاقد داده‌های پرت یا در حالت خوش‌بینانه فاقد عیب است که برای مدل‌سازی منحنی توان یا آموزش مدل‌های یادگیری ماشین استفاده گردد. (عبارت خوش‌بینانه به‌این‌علت مطرح شد که به دلیل عدم وجود برچسب، صحت سنجی ممکن نیست و هنگام کار با داده‌های نابهنجار، عملاً برچسبی وجود ندارد و ماهیت ذاتی نابهنجاری نیز بر پایه‌ی داده‌ی بدون برچسب استوار است. به‌محض وجود برچسب، کار تشخیص نابهنجاری فاقد اعتبار خواهد بود و مستقیماً امکان تشخیص عیب وجود خواهد داشت.)

۲- الگوریتمی ساده، دقیق و بدون وابستگی شدید به ساختار توربین بادی برای این منظور توسعه‌یافته است.

ساختار باقی مقاله در ادامه مطرح می‌شود. بخش دوم مربوط به منحنی توان و مفاهیم کلی خواهد بود که پیش‌نیازهای ضروری برای کار را مطرح می‌کند. روش‌های پیشنهادی در بخش سوم معرفی می‌شود. درنهایت نتایج تجربی در بخش چهارم مورد بحث قرار می‌گیرد.

۲- مفاهیم موردنیاز: منحنی توان، دی‌بی‌اسکن، مدل یادگیری و آموزش، شاخص‌های ارزیابی

الف) منحنی توان

منحنی توان توربین بادی (Power Curve) یکی از شاخص‌های اصلی در ارزیابی و تحلیل کارایی سامانه‌های انرژی بادی محسوب می‌شود [۲۶] (شکل ۱). این منحنی رابطه‌ی غیرخطی میان سرعت باد در ارتفاع هاب و توان الکتریکی خروجی توربین را نشان می‌دهد و از همین رو، مبنای مهمی در مدل‌سازی رفتار دینامیکی سامانه به شمار می‌رود. منحنی توان اطلاعات ارزشمندی در زمینه‌هایی چون برآورد انرژی تولیدی و شناسایی داده‌ی نابهنجاری‌ها در اختیار قرار می‌دهد. توان تولیدی یک توربین بادی به‌صورت رابطه‌ی (۱) مدل‌سازی می‌شود:

$$p = \frac{1}{2}\rho A C_P(\lambda,\beta) v^3 \qquad (۱)$$

که در آن توان خروجی توربین (وات)، ρ چگالی هوا بر حسب (kg/m^3)، A مساحت جاروب شده توسط پره‌ها، C_P ضریب توان توربین و تابعی از نسبت سرعت نوک پره λ و زاویه گام پره β، v سرعت باد در ارتفاع هاب (متر بر ثانیه) است.

ناحیه‌های عملکردی توربین بادی به چهار بخش تقسیم می‌شود که در شکل ۱ نشان‌داده‌شده است و به شرح زیر است:

- ۱- ناحیه اول ($v < v_{\text{cut-in}}$):

در این بازه، به دلیل پایین‌بودن سرعت باد، توربین هیچ توانی تولید نمی‌کند.

- ۲- ناحیه دوم ($v_{\text{cut-in}} \le v < v_{\text{rated}}$):

با افزایش سرعت باد از سرعت قطع ورود (cut-in) فرایند تولید انرژی آغاز شده و تا سرعت نامی (rated)، توان خروجی به طور غیرخطی افزایش می‌یابد.

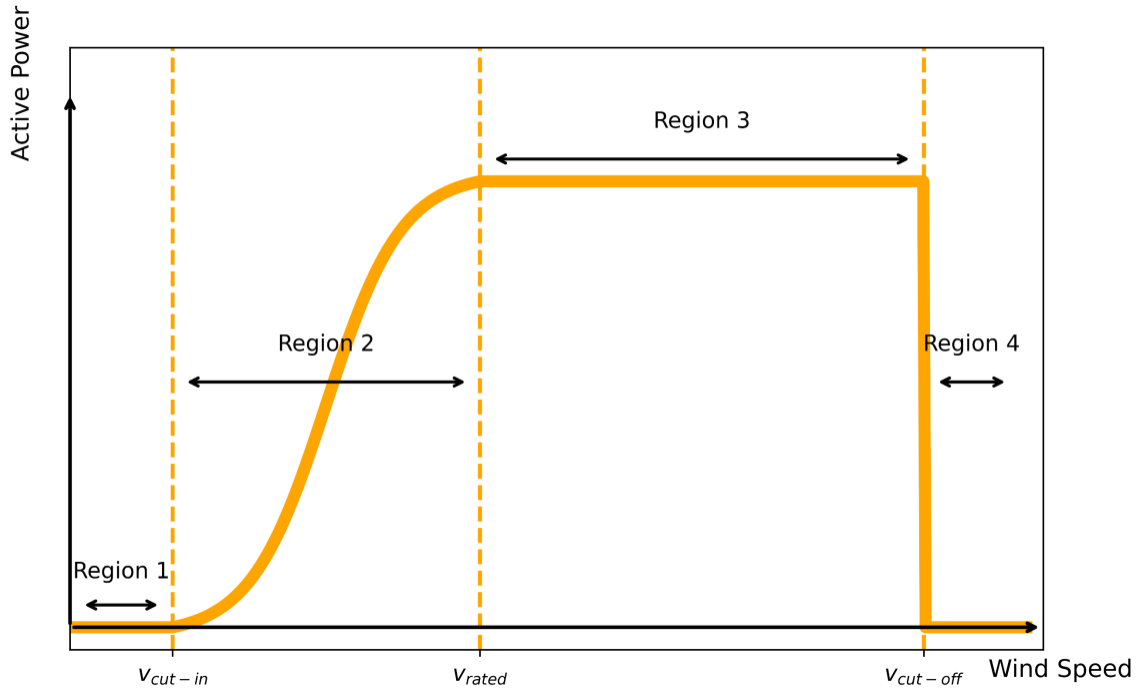
- ۳- ناحیه سوم ($v_{\text{rated}} \le v < v_{\text{cut-off}}$):

در این بازه، توربین، توان نامی خود را به طور ثابت و پایدار تولید می‌کند.

- ۳- ناحیه چهارم ($v \ge v_{\text{cut-off}}$):

در سرعت‌های بسیار بالا، برای جلوگیری از آسیب‌دیدن اجزای مکانیکی، توربین به‌صورت خودکار متوقف می‌شود.

ازآنجاکه چگالی هوا در مکان‌ها و شرایط آب‌وهوایی مختلف متغیر است



شکل ۱. منحنی ایدئال توربین بادی.

Fig. 1. Ideal power curve of a wind turbine

جهت اصلاح سرعت باد اندازه‌گیری‌شده از رابطه‌ی (۲) استفاده می‌شود که به آن اصلاح چگالی هوا گفته می‌شود:

$$v_C = v_M \left(\frac{\rho}{\rho_0} \right)^{\frac{1}{3}} \quad (۲)$$

دراین‌رابطه سرعت باد اصلاح‌شده،سرعت باد اندازه‌گیری‌شده، چگالی واقعی هوا، چگالی استاندارد در سطح دریا است.

به‌منظور رسم منحنی توان واقعی توربین، معمولاً از روش بازه‌بندی استفاده می‌شود. در این روش داده‌ها بر اساس بازه‌های سرعت باد دسته‌بندی شده و میانگین توان برای هر بازه محاسبه می‌شود:

$$P_i = \frac{1}{n_i} \sum_{j=1}^{n_i} P_{ij} \quad (۳)$$

که در آن P_i میانگین توان در بازه شماره i، P_{ij} مقدار توان برای نمونه‌ی j در همان بازه، و n_i تعداد کل داده‌های آن بازه است.

شایان‌ذکر است که منحنی توان ایدئال ارائه شده توسط سازندگان بر اساس استانداردهایی مانند IEC است و شرایط واقعی مانند تغییرات محیطی و فرسایشی را در نظر نمی‌گیرد. برعکس، منحنی توان حاصل از داده‌های میدانی مانند اسکادا، بازتاب دقیقی از عملکرد واقعی توربین ارائه می‌دهد و به همین علت، در کاربردهایی چون پایش وضعیت اهمیت بیشتری پیدا می‌کند [۳].

روش‌های مختلف پارامتریک[1] و غیرپارامتریکی[2] برای مدل‌سازی منحنی توان توربین بادی وجود دارد که به‌صورت مفصل در [۲۷, ۲۸] اشاره شده است. به طور مثال، از مدل‌های پارامتری می‌توان به مدل تکه‌ای‌خطی[3]، چندجمله‌ای[4] و از مدل‌های غیرپارامتری به شبکه‌های عصبی[5]، مدل‌های یادگیری ماشین، مدل‌های فازی اشاره نمود. به طور خلاصه، از بین این روش‌ها، رویکردهای مبتنی بر یادگیری ماشین بر رویکردهای دیگر برتری

1. Parametric
2. Non-parametric
3. Linearized Segmented model
4. Polynomial
5. Neural Networks

دارد؛ چراکه با پرهیز از ساده‌سازی‌های غیرواقعی، به تقریب بهتری از سامانه واقعی می‌رسد. علاوه بر این، نیاز به شناخت روابط فیزیکی حاکم بر مسئله نیست و به‌اصطلاح، مانند جعبه سیاه عمل می‌کنند.

ب) دی‌بی‌اسکن

برای تشخیص داده‌ی نابهنجار و یا عیب با استفاده از مدل‌های یادگیری ماشین، قبل از هر کاری باید داده‌های خام، پاک‌سازی شود و داده‌های پرت از مجموعه‌داده حذف شوند. نقاط پرت بر روی دقت مدل تأثیر منفی می‌گذارند و مدل را به سمت یادگیری الگوهای اشتباه سوق می‌دهند. یکی از الگوریتم‌های معروف در این زمینه DBSCAN[1] است. روش کار این الگوریتم بر پایه‌ی چگالی محلی داده‌ها استوار است و برخلاف روش‌های مرسوم، نیازی به تعیین تعداد خوشه‌ها در آغاز کار ندارد. در دی‌بی‌اسکن، اگر چگالی داده‌ها اطراف داده‌ی مدنظر، کم‌تر از مقدار مطلوب باشد آن داده به‌عنوان داده‌ی دورافتاده و پرت در نظر گرفته می‌شود. میزان همسایگی داده با شعاع ε نشان داده می‌شود و معیار دوری با استفاده از تابع فاصله (مانند فاصله اقلیدسی) تعریف می‌شود. این بررسی برای تک‌تک نقاط مجموعه‌داده تکرار می‌شود تا همسایگی همه‌ی نقاط بررسی و همه‌ی نقاط پرت شناسایی شود. روابط دی‌بی‌اسکن در ادامه نشان‌داده‌شده است.

$$N_{\varepsilon}(p)=\{q\in \mathrm{D} \mid \mathrm{dist}(p,q)\le \varepsilon\} \quad (۴)$$

$$\mathrm{dist}(p,q)=\sqrt{\sum_{i=1}^{d}(p_i-q_i)^2} \quad (۵)$$

$$\left|N_{\varepsilon}(p)\right|\ge \mathrm{Minpts} \quad (۶)$$

مجموعه‌ی همسایگی یا در (۴) مشخص شده است و به مجموعه نقاط q گفته می‌شود که در شعاع همسایگی از نقطه‌ی p باشد. معیار فاصله در (۵) تعریف شده است که همان فاصله‌ی اقلیدسی است (d به معنای بُعد است). به نقطه‌ی p که بیشتر از تعدادی مشخص داده (Minpts) در اطراف آن باشد هسته گفته می‌شود (در رابطه‌ی (۶) صدق کند).

1. Density-Based Spatial Clustering of Applications with Noise

همان‌طور که در مقدمه نیز اشاره شد در توربین بادی و پیش از اعمال این الگوریتم باید داده‌های خطامانند را به‌وسیله‌ی یک فیلتر اولیه حذف کرد. اعمال فیلتر، موجب سوگیری به نفع روش پیشنهادی در این پژوهش نمی‌شود؛ بلکه برعکس، شرط لازم برای عملکرد قابل‌قبول دی‌بی‌اسکن است؛ در نتیجه، بدون فیلتر عملاً پاک‌سازی انجام نمی‌شود و دقت مدل آموزش‌دیده بر پایه‌ی دی‌بی‌اسکن کاهش می‌یابد. به همین دلیل، دی‌بی‌اسکن بدون فیلتر در همان مرحله‌ی اولیه ناکام می‌ماند و تنها پس از اعمال فیلتر است که این الگوریتم امکان رقابت با روش پیشنهادی را پیدا می‌کند؛ درحالی‌که روش معرفی شده ذاتاً بدون نیاز به فیلتر عملکرد مؤثری دارد؛ بنابراین، ساختار مقایسه ناعادلانه نیست و سوگیری ایجاد نمی‌شود.

پ) مدل یادگیری و آموزش

پس از پیش‌پردازش داده‌های خام، انتخاب ویژگی و آموزش مدل قرار انجام می‌شود.

در انتخاب ویژگی، به‌جای استفاده از کل ویژگی‌ها به‌عنوان ورودی که باعث افزایش هزینه‌ی محاسباتی می‌شود با انتخاب ویژگی مناسب، علاوه بر کاهش هزینه‌ی محاسباتی، افزایش دقت مدل نیز ممکن خواهد بود. در این پژوهش با استفاده از ماتریس همبستگی، ویژگی‌های مرتبط با توان خروجی، استخراج می‌شود. در این پژوهش از سرعت باد به‌عنوان ورودی استفاده شده است (در واقع، تشخیص عیب مبتنی بر منحنی توان خواهد بود).

در گام بعدی مدل آموزش می‌بینند. در این پژوهش از الگوریتم K نزدیک‌ترین همسایه (KNN)[2] استفاده شده است که یکی از ساده‌ترین و درعین‌حال مؤثرترین روش‌های یادگیری ماشین محسوب می‌شود. برخلاف سایر الگوریتم‌های آموزش، در این روش، رابطه‌ی ریاضی صریحی جهت پیش‌بینی ایجاد نمی‌شود؛ بلکه مدل، تمام داده‌های آموزشی را ذخیره و تصمیم‌گیری را به زمان پیش‌بینی موکول می‌نماید. درواقع، این روش با بررسی حلقه‌ای از نزدیک‌ترین همسایه‌های داده‌ی ورودی، مقدار خروجی را تخمین می‌زند. مبنای کار این روش این نکته است که نمونه‌های ورودی مشابه، امکان بالاتری برای داشتن مقادیر هدف مشابه دارند. مزیت این الگوریتم، سادگی مفهومی و قابلیت مدل‌سازی روابط پیچیده و غیرخطی است که کاربرد گسترده‌ای در مسائل رگرسیون و طبقه‌بندی دارد.

نکته‌ی بعدی هم‌تراز کردن ورودی‌ها در استفاده از الگوریتم K نزدیک‌ترین همسایه است؛ چراکه در صورت وجود چندین ورودی مختلف

2. K-Nearest Neighbors

با مقیاس‌های متفاوت، اثر برخی از ورودی‌ها به‌کلی نادیده گرفته می‌شود و مدل برای تشخیص به برخی ویژگی‌ها ناعادلانه و نادرست، بیشتر اهمیت می‌دهد. در این صورت عملاً کار یادگیری با مشکل مواجه می‌شود و سبب یادگیری ناموثر مدل شود. برای حل این مشکل، استانداردسازی یا نرمال‌سازی داده‌ها به کار گرفته می‌شود. در این پژوهش از نرمال‌سازی به‌صورت رابطه‌ی (۷) استفاده شده است.

$$\frac{x-\mu}{\sigma}=z \qquad (۷)$$

که در آن x مجموعه ویژگی‌ها، μ میانگین، σ انحراف معیار ویژگی‌ها و مجموعه‌ی ویژگی‌ها است که دارای توزیع نرمال خواهد بود. با این کار، تمامی ویژگی‌ها در مقیاس‌های منطقی قرار می‌گیرند و تأثیر فواصل معنادار آن‌ها بر مدل، خنثی می‌شود.

الگوریتم K نزدیک‌ترین همسایه در جدول ۱ آمده است و همان‌طور که اشاره شد مدل با مراجعه به نزدیک‌ترین همسایه‌ها در فضای ویژگی، مقدار پیش‌بینی‌شده را تخمین می‌زند.

در جدول ۱، x ورودی i–ام، y_i مقدار خروجی i–ام، $\hat{y}$ مقدار پیش‌بینی خروجی و n تعداد داده‌ها است.

ت) شاخص‌های ارزیابی

پس از آموزش مدل‌های یادگیری ماشین و تخمین منحنی توان، عملکرد مدل با بهره‌گیری از مجموعه‌داده‌ی تست و شاخص‌های خطای متداول مورد ارزیابی قرار می‌گیرد. چهار شاخص ارزیابی معروف و پرکاربرد در این پژوهش به کار می‌رود که شامل میانگین قدرمطلق خطا (MAE)، میانگین مربعات خطا (MSE)، ریشه میانگین مربعات خطا (RMSE) و ضریب تعیین (R^2) هستند که روابط آن‌ها در (۸) تا (۱۰) آمده است.

$$\mathrm{MAE}=\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\left|y_i-y_i\right| \qquad (۸)$$

$$\mathrm{MSE}=\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\left(y_i-y_i\right)^2 \qquad (۹)$$

جدول ۱. الگوریتم K نزدیک‌ترین همسایه.

Table 1. KNN algorithm.

ورودی:
مجموعه‌داده‌های $\mathbf{D}=\{(x_i,y_i)\}_{i=1}^{n}$ **(تمام جفت‌های ورودی و خروجی)، نمونه جدید x، پارامتر K**
الگوریتم:
محاسبه فاصله (تشابه) نقطه‌ی x تا تمام نمونه‌های آموزش:
$i=1,\dots,n$ **برای** $d_i=distance(x,x_i)$
مرتب‌سازی فاصله‌ها به‌صورت صعودی
انتخاب K نزدیک‌ترین همسایه
محاسبه پیش‌بینی:
میانگین ساده: $\hat{y}=\left(\frac{1}{K}\right)\sum_{i=1}^{k}y_i$
خروجی:
مقدار پیش‌بینی شده $\hat{y}$

$$\mathrm{RMSE}=\sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{\mathrm{n}}\left(y_i-y_i\right)^2} \qquad (۱۰)$$

در روابط (۸) تا (۱۰)، n تعداد کل داده‌هایی که برای سنجش مدل به کار می‌رود،

y_i مقدار واقعی داده مربوط به مشاهده‌ی i–ام، $\hat{y}$ مقدار پیش‌بینی‌شده توسط مدل برای همان مشاهده، و $\bar{y}$ میانگین مقادیر واقعی است.

MAE نشان‌دهنده‌ی میانگین قدرمطلق اختلاف بین مقادیر واقعی و پیش‌بینی‌شده است و مقدار کمتر آن بیانگر عملکرد بهتر مدل است. MSE میانگین مربع اختلاف مقادیر واقعی و برآوردی را محاسبه می‌کند و به دلیل مجذور کردن خطاها، حساسیت بیشتری به خطاهای بزرگ‌تر دارد و همانند MAE، مقدار کمتر نشان‌دهنده‌ی عملکرد بهتر است. همانند MSE است، با این تفاوت که با گرفتن ریشه دوم، واحد آن با داده‌های اصلی سازگار می‌گردد و تفسیرپذیری بهتری دارد. مقدار کمتر RMSE نشانه عملکرد بهتر مدل است. R^2 یا ضریب تعیین میزان واریانس توضیح داده شده توسط مدل را نشان می‌دهد. این شاخص در بازه ۰ تا ۱ قرار می‌گیرد و هرچه به عدد ۱ نزدیک‌تر باشد بیانگر دقت بالاتر مدل در برازش داده‌های واقعی است.

۳- روش پیشنهادی

استفاده از چندک‌ها در دو گام پیشنهاد می‌شود: ۱- انتخاب نقاط نماینده ۲- حدود آستانه.

در گام اول، از داده‌های نماینده استفاده می‌شود [۲۹]. یکی از روش‌های ساده و کاربردی، تقسیم محدوده‌ی موردنظر به بازه‌های مساوی و سپس میانگین‌گیری و انتخاب میانگین‌ها به‌عنوان داده‌های نماینده است. ازآنجاکه داده‌های پرت تأثیر قابل‌توجهی بر میانگین می‌گذارند (به‌ویژه در بازه‌های خاصی از سرعت باد مانند شرایط عملکرد کمتر از حد انتظار یا خاموشی حفاظتی توربین که توان خروجی با افت ناگهانی یا نویز شدید همراه است) اتکا به میانگین ساده، موجب انحراف نتایج از مقادیر واقعی توان و نهایتاً سوگیری مدل می‌شود. همچنین، در داده‌های اسکادای منحنی توان این مقادیر پرت عمدتاً به سمت پایین منحنی توان مشاهده می‌شوند؛ برای مثال در شرایط توان صفر، کاهش اجباری تولید، توقف ناگهانی یا کارکرد در حالت توان کاهش‌یافته که همگی میانگین را به طور ساختارمند کمتر از سطح واقعی عملکرد توربین نشان می‌دهند. برای تقسیم‌بندی سرعت باد و تعیین تعداد بازه‌ها، روش‌های معروف گوناگونی وجود دارد که از آن‌ها می‌توان به روش قانون استورجس (صرفاً مبتنی بر تعداد داده و مستقل از ساختار داخلی توربین) یا استانداردهای IEC (بازه‌بندی به‌صورت ۰/۵ متر بر ثانیه برای سرعت باد) اشاره نمود. در این پژوهش، مبنای بازه‌بندی، استاندار IEC است به‌طوری که جستجوی شبکه‌ای در حول مقدار به‌دست‌آمده از IEC صورت می‌گیرد تا بهترین تعداد بازه با کمترین خطاها و سرعت مناسب انتخاب شود. شاخص ارزیابی در جستجوی شبکه‌ای از سبک‌وسنگین کردن MAE (به‌عنوان نماینده از خطا) و بهترین زمان به دست می‌آید. همچنین می‌توان معیارهای ترکیبی و مبتنی بر خواست متخصص بهره‌برداری به‌صورت سفارشی تنظیم نمود (مثلاً شاخصی از مجموعه‌ی معیارها با وزن‌های متفاوت برای هر یک بر اساس نوع نیاز).

برخلاف استفاده از میانگین برای نمایندگی داده‌ها که تأثیر زیادی از داده‌های پرت می‌پذیرد بهره‌گیری از شاخص‌های مقاوم‌تر مانند چندک‌ها امکان تمرکز بر بخشی معین از توزیع داده‌ها را فراهم می‌کند (به‌ویژه مقادیر بالاتر در منحنی توان که بیانگر عملکرد پایدار و بهینه توربین هستند) و چندک‌ها می‌توانند بازتابی دقیق‌تر از حداکثر توان بالقوه و شرایط واقعی عملکرد توربین ارائه می‌دهند. پس از انتخاب سرعت باد به‌عنوان ویژگی مناسب با استفاده از ماتریس همبستگی، امکان پیاده‌سازی چندک و انتخاب نقاط نماینده فراهم می‌شود. ابتدا داده‌های سرعت باد به بازه‌هایی با تعداد نمونه‌های تقریباً مساوی تقسیم می‌شوند تا در هر بازه، رفتار منحنی توان به طور یکنواخت بررسی و ناهمگونی در توزیع منحنی توان در نظر گرفته شود. سپس برای اطمینان بیشتر، تنها بازه‌هایی که تعداد نمونه کافی دارند (بیش از یک آستانه تعیین‌شده باشند) نگه داشته می‌شوند تا از ناپایداری آماری جلوگیری شود. در نهایت، برای هر بازه‌ی معتبر، چندک دلخواه از توان خروجی و سرعت باد متناظر انتخاب می‌شود. نتیجه این فرایند، تولید منحنی توانی هموارتر با داده‌های کمتر است که در آن اثر نقاط پرت و نوسانات تصادفی به طور محسوسی کاهش‌یافته است؛ ولی رفتار کلی توربین در نظر گرفته شده است. در این پژوهش، مقدار پیشنهادی برای چندک در گام پیش‌پردازش، چندک ۵۰ ام یا میانه است. مزیت این انتخاب در این است که علاوه بر کاهش تأثیر نویز و نقاط پرت، مقدار میانه می‌تواند برای ساختارهای متفاوت توربین بادی ثابت بماند و نیازی به تنظیم دقیق و شدید، بسته به ساختار داده و توربین بادی ندارد و از پیچیدگی انتخاب آن جلوگیری می‌شود.

سپس مدل با داده‌های پاک‌سازی شده آموزش می‌بیند و از مقایسه‌ی

مقدار تخمین لحظه‌ای مدل با داده‌های لحظه‌ای ثبت‌شده توسط سامانه اسکادا، سیگنالی موسوم به مانده (residual) محاسبه می‌شود (رابطه‌ی (۱۱)) که میزان انحراف از این سیگنال، بیانگر وقوع نابهنجاری و انحراف از شرایط مطلوب است. در صورت عبور مقدار مانده از یک آستانه‌ی تعیین‌شده (رابطه‌ی (۱۲))، رفتار غیرعادی (و نه وقوع یک عیب) شناسایی و گزارش می‌شود؛ بنابراین تحلیل سیگنال باقی‌مانده، امکان شناسایی نقاط پرت در مرحله‌ی پیش‌پردازش و داده‌ی نابهنجار بعد از آموزش مدل را فراهم می‌کند.

$$r_t = x_t - \text{Predicted value} \tag{۱۱}$$

$$|r_t| \geq 3\times\sigma_r \tag{۱۲}$$

در روابط (۱۱) و (۱۲)، r_t مقدار باقی‌مانده در زمان t، x_t مقدار داده‌ی واقعی در آن لحظه و σ_r مقدار انحراف معیار سیگنال باقی‌مانده است.

پس از آموزش مدل و در گام بهره‌برداری و تشخیص، دوباره استفاده از حدود آستانه‌ی چندکی پیشنهاد می‌شود. معمولاً در تعیین حدود آستانه و برای ساده‌سازی، سیگنال مانده در توربین به‌صورت توزیع نرمال فرض می‌شود و از حد آستانه روش ۳-سیگما استفاده می‌شود. در این روش، بازه‌ی اطمینان به‌صورت $[\mu + ۳\sigma \ , \mu - ۳\sigma]$ تعریف می‌شود که در آن μ میانگین سیگنال مانده و σ انحراف معیار آن است. درصورتی‌که رفتار سامانه به‌صورت نرمال باشد بیش از ۹۹٫۷٪ داده‌ها در این بازه قرار می‌گیرند؛ بنابراین، مقادیر مانده‌ای که از این بازه (رابطه‌ی (۱۲)) خارج شوند با احتمال بسیار بالا، نقطه‌ی غیرعادی یا نابهنجار است.

بااین‌حال، اگر عمیق‌تر نگاه شود سیگنال مانده در توربین بادی واقعاً و کاملاً به‌صورت نرمال نیست. در برخی از موارد، الگوی توزیع داده‌ها ممکن است دارای کشیدگی دم یا به‌صورت کلی چولگی باشند که باعث می‌شود استفاده از روش ۳-سیگما منجر به شناسایی نادرست شود؛ چراکه اساس کار آن، بر ناحیه‌ی اطمینان و فرض توزیع نرمال استوار است.

بنابراین، برای تحلیل واقع‌بینانه‌تر و دقیق‌تر رفتار داده‌ها، استفاده از چندک محلی پیشنهاد می‌شود. چندک یا q به طور ساده مقداری است که یک درصد مشخصی از داده‌ها از نظر مقداری، زیر آن قرار دارند. به‌عنوان‌مثال، ۹۵ امین چندک (یا q_{95}) مقداری است که ۹۵٪ داده‌ها از نظر مقداری کمتر از آن هستند و تنها ۵٪ داده‌ها بالاتر از آن قرار می‌گیرند. در این رویکرد، به‌جای فرض نرمال بودن توزیع و استفاده از ۳-سیگما، با بازه‌بندی سیگنال، حدود آستانه مبتنی بر توزیع واقعی داده‌ها تعیین می‌شود و در نتیجه، کارایی مبتنی بر واقعیت از تشخیص نقاط پرت ارائه می‌دهد؛ بنابراین به‌جای استفاده از (۱۲) حد آستانه‌ی (۱۳) و (۱۴) پیشنهاد می‌شود.

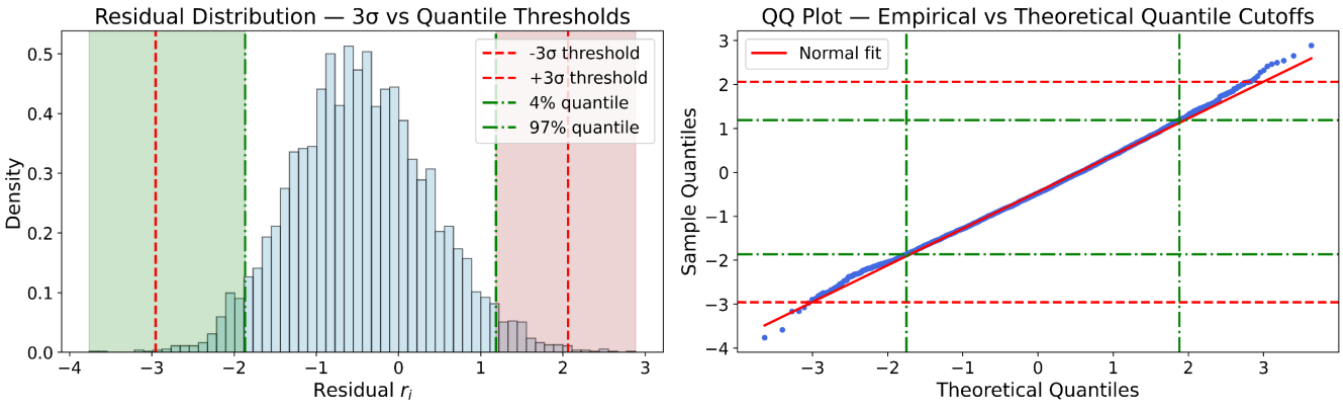
$$|r_t| \succ \mathrm{Q_p} \tag{۱۳}$$

$$|r_{\mathrm{Rt}}| \succ \mathrm{Q_{Rp,}} \quad |r_{\mathrm{Lt}}| \succ \mathrm{Q_{Lp,}} \tag{۱۴}$$

که در آن r_{Rt} بخش مثبت سیگنال مانده یا دم راست، r_{Lt} بخش منفی سیگنال مانده یا دم چپ و $\mathrm{Q_{Rp}}$ و $\mathrm{Q_{Lp}}$ به ترتیب مقدار چندک برای حد آستانه‌ی طرف چپ و راست است.

با تنظیم Qpها به‌عنوان حد آستانه و مرز حذف داده‌ها، داده‌هایی که سیگنال مانده‌ای غیرمتعارف از خود نشان دهند به‌عنوان نقاط نابهنجار در نظر گرفته و حذف می‌شوند. برای بالابردن دقت و درنظرگرفتن نامتقارنی توزیع سیگنال در رفتار توربین بادی، در صورت وجود چولگی و باتوجه‌به نامتقارن بودن توزیع، به نیمه‌های مثبت و منفی سیگنال رفتار متفاوتی اعمال می‌شود این رویکرد در شکل ۲ مشخص است.

این شکل نشان می‌دهد روش آستانه‌ی چندکی با داشتن آزادی عمل بیشتر نسبت به روش ۳-سیگما، داده‌های نابهنجار را به‌صورت دقیق‌تر شناسایی می‌کند. در واقع با تطبیق حدود آستانه بر شروع انحراف در سیگنال باقی‌مانده می‌توان به طرز واقع‌بینانه‌تری داده‌ها را شناسایی و گزارش نمود. مقدار مناسب این حدود آستانه‌ی چندکی، بر اساس سیگنال باقی‌مانده یا روش‌هایی مانند تحلیل منحنی QQ به دست می‌آید؛ اما باتوجه‌به رفتار نزدیک به نرمال سیگنال باقی‌مانده در توربین بادی می‌توان برای بسیاری از توربین‌های بادی مقادیر متداول مانند ۹۰ و ۱۰ را برای آستانه‌ی بالا و پایین (راست و چپ) جایگزین نمود بدون اینکه به‌دقت مدل‌سازی لطمه‌ی جدی بخورد که در بخش نتایج نیز نشان‌داده‌شده است. در صورت عدم فرض نزدیک به نرمال بودن، معمولاً در داده‌های توربین بادی باکیفیت، مقادیر مناسب با اندکی آزمون‌وخطا حول نقاط ۹۰ و ۱۰، به دست می‌آیند؛ چراکه در این‌گونه داده‌ها، درصد کمی نویز و نقاط پرت وجود دارد.



شکل ۲. توزیع شبه نرمال و حدود آستانه.

Fig. 2. Quasi normal distribution and threshold bounds.

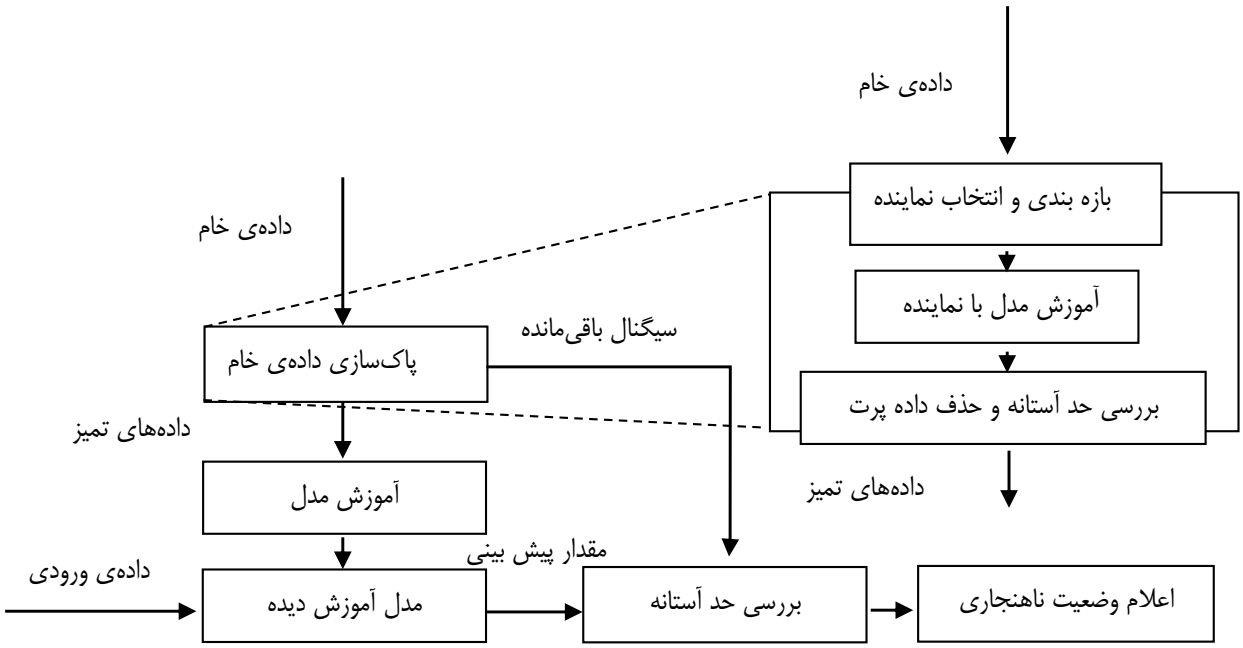
جدول ۲. الگوریتم تشخیص دادهی نابهنجار.

Table 2. Abnormal data detection algorithm.

ورودی:
D $= \{(x_i, y_i)\}_{i=1}^{n}$ **(مجموعهدادهی تمیز شده)**
الگوریتم:
۱. پیشپردازش و پاکسازی دادههای خام
۲. آموزش مدل یادگیری
۲. تعیین حد آستانه بر اساس چندک و سیگنال باقیمانده، رابطهی (Error! Reference source not found. و Error! Reference source not found.)
۳. بررسی وضعیت دادهی ورودی بر اساس حد آستانه
خروجی:
اعلام وضعیت داده

پس از تعیین حد آستانه جهت تشخیص نابهنجاری (برای نمونه ۹۰ و ۱۰)، با استفاده از دادههای تمیز، مدل K نزدیکترین همسایه آموزش میبیند و باتوجهبه حد آستانهی (۱۳) به تشخیص نقاط دادهی نابهنجار میپردازد بهطوری که چنانچه دادهای خارج از این حد آستانه باشد آنگاه بهعنوان دادهی نابهنجار (دادهای با رفتار غیرعادی) اعلام میشود.

به طور خلاصه، ابتدا دادهها پاکسازی میشود و سیگنال باقیمانده ذخیره میشود. این پاکسازی با استفاده از بازهبندی و حد آستانهی چندکی اجرا میگردد. سپس مدل یادگیری ماشین با این دادههای تمیز آموزش میبیند و باتوجهبه سیگنال باقیمانده و حد آستانهی چندکی محلی (در هر بازه جداگانه)، وضعیت نابهنجار بودن بررسی و اعلام میشود. الگوریتم و بلوک دیاگرام کلی تشخیص دادهی نابهنجاری به ترتیب در جدول ۲ و شکل ۳ آمده است.



شکل ۳. بلوک دیاگرام تشخیص داده‌ی نابهنجاری.

Fig. 3. Abnormal data detection block diagram.

۴- آزمایش و نتایج

آزمایش پیش رو بر اساس داده‌های اسکادای یک مزرعه توربین بادی واقعی واقع در ترکیه صورت‌گرفته است. داده‌ها در بازه‌های ۱۰ دقیقه‌ای و شامل حدود ۵۰ هزار داده برای سال ۲۰۱۸ است که طی مدت‌زمان یک سال ثبت شده است. خلاصه‌ای از ساختار این مجموعه‌داده در جدول ۳ آمده است.

منحنی توان توربین بادی این داده‌ها در شکل ۴ نشان‌داده‌شده است. نقاط پرت در قالب ضربدر و به رنگ قرمز در نمودار مشخص است که لازم است در مرحله‌ی پیش‌پردازش و قبل از آموزش مدل، حذف شوند.

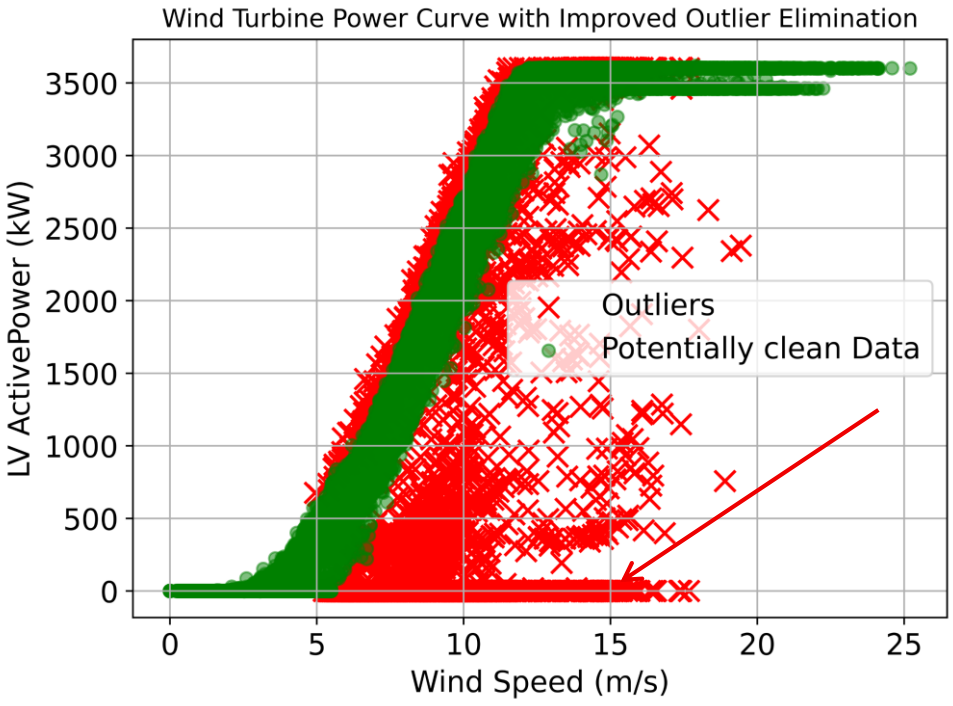
روش پاک‌سازی اول دی‌بی‌اسکن است؛ اما این الگوریتم در شناسایی برخی نقاط پرت محدودیت دارد. همان‌طور که در شکل ۴ مشاهده می‌شود این الگوریتم در تشخیص داده‌های پرت متراکم و خطامانند در محدوده سرعت باد ۵ تا حدود ۱۷ متر بر ثانیه ناموفق عمل می‌کند و قادر به حذف آن‌ها نیست (داده‌هایی که فلش به آن‌ها اشاره می‌کند در شکل ۴ و شکل ۵).

در واقع، دی‌بی‌اسکن نه‌تنها قادر به شناسایی و حذف این بخش از داده‌ها به‌عنوان نقاط پرت نیست؛ بلکه آن‌ها را به‌عنوان داده‌های معتبر طبقه‌بندی می‌نماید. این در حالی است که بر اساس مطالعات پیشین (که به برخی در مقدمه پژوهش اشاره شد) و ملاحظات نظری، این ناحیه از عملکرد توربین بادی که مربوط به وضعیت غیرایده‌آل و گذراست تأثیر منفی بر دقت مدل‌سازی دارد و باید از مجموعه‌داده حذف گردد. این ضعف الگوریتم ناشی از ماهیت پارامترهای دی‌بی‌اسکن و تراکم بالای نقاط در این محدوده است که منجر به تشکیل یک خوشه ظاهراً معتبر می‌شود؛ بنابراین الگوریتم دی‌بی‌اسکن در همین ابتدای کار نیاز دخالت و حذف با فیلتری جداگانه برای نقاط خطامانند دارد وگرنه عملکرد تخمین مدل به‌شدت تحت‌تأثیر این نقاط پرت حذف نشده قرار می‌گیرد و دقت آن کاهش می‌یابد. پس با تعریف فیلتری با دو محدوده در محورهای سرعت باد و توان خروجی، این نقاط از مجموعه‌داده حذف می‌شود. این محدوده‌ها به‌ازای توربین بادی‌های مختلف، متفاوت است و باید بر اساس ساختار هر توربین، جداگانه و از نو تعیین شوند. پس از حذف به‌وسیله‌ی فیلتر، باقی‌مانده‌ی داده‌ها جهت پاک‌سازی به دی‌بی‌اسکن داده می‌شود و در نهایت، داده‌های پاک‌سازی‌شده‌ی خروجی، برای مدل‌سازی به مدل داده می‌شود. پارامترهای الگوریتم می‌تواند تصادفی یا تجربی انتخاب شود که بر دقت الگوریتم شدیداً تأثیر می‌گذارد. در این پژوهش انتخاب بر اساس جستجوی شبکه‌ای برای تعدادی محدود از مقادیر (نقاط ۳، ۴، ۵، ۶ و شعاع ۰/۱، ۰/۲،...، ۰/۶) و مبتنی بر شاخص سیلوئت انتخاب شده است. این شاخص میزان فشردگی و جدایی خوشه‌ها را می‌سنجد

جدول ۳. ساختار داده‌های بدون برچسب

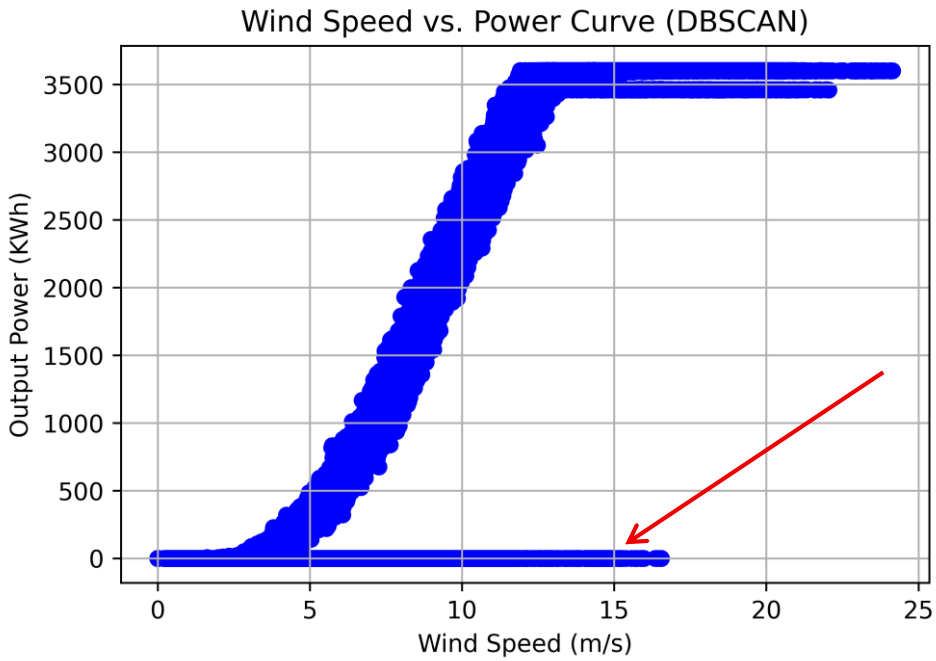
Table 3. Structure of unlabeled dataset.

ستون	توضیحات
Date/Time	زمان و تاریخ ثبت داده‌ها در بازه‌های ۱۰ دقیقه‌ای
LV ActivePower (kW)	توان لحظه‌ای تولید شده توسط توربین (کیلووات)
Wind Speed (m/s)	سرعت باد در ارتفاع هاب توربین (متر بر ثانیه)
Theoretical_Power_Curve (kW)	توانی نامشخص تحت عنوان تئوری (کیلووات)
Wind Direction (°)	جهت باد در ارتفاع هاب توربین (درجه)



شکل ۴. منحنی توان توربین بادی.

Fig. 4. Wind turbine power curve.



شکل ۵. منحنی توان پاک‌سازی شده توربین بادی با دی‌بی‌اسکن بدون استفاده از فیلتر.

Fig. 5. Cleaned wind turbine power curve using DBSCAN without filtering.

و مقداری بین ۱− (ضعیف) و ۱+ (عالی) دارد و مقادیر مناسب برای ساختار الگوریتم را می‌توان از این طریق به دست آورد. ساختار دی‌بی‌اسکن استفاده شده و ساختار فیلتر تعریف شده در جدول ۴ و نمودار پاک‌سازی شده در شکل ۶ آمده است.

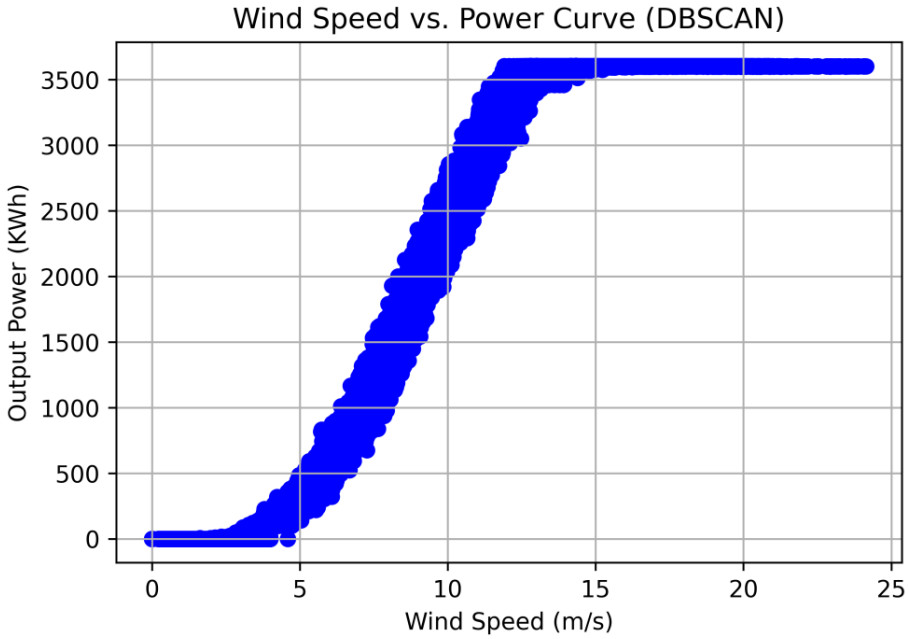
روش دوم پاک‌سازی، روش پیشنهادی است. این روش بدون نیاز به اعمال فیلتر برای هر توربین، صرفاً با تعیین تعداد بازه‌ها و چندک، پاک‌سازی را انجام می‌دهد. مزیت این رویکرد این است که این پارامترها معمولاً برای بیشتر توربین‌های بادی ثابت نگه داشته می‌شوند و نیازی به تنظیم‌های مکرر و توربین به توربین ندارند. همان‌طور که قبلاً اشاره شد استفاده از چندک میانه (۵۰) برای انتخاب داده‌ی نماینده از هر بازه، رفتار معمول توربین را بدون اثرگذاری داده‌های پرت استخراج می‌کند و برای اغلب توربین‌ها مقدار مناسب و قابل‌اعتماد است. تعداد بازه‌ها نیز طبق جستجو حول مقدار ارائه شده توسط IEC یعنی ۰/۵ متر بر ثانیه و توسط آزمون‌وخطا انجام می‌شود تا مقدار مناسب استخراج شود. بر طبق این استاندارد، باتوجه‌به سرعت باد باید حدود ۵۰ بازه در نظر گرفت. برای دقت بهتر و تعداد بازه‌ی متناسب با داده‌ها، آزمون‌وخطا حول این مقدار انجام می‌شود و تعداد بازه برای مقادیر ۱۰، ۲۰، و... تا ۱۵۰ بررسی می‌شود و سپس بهترین مقدار بر اساس مقدار خطای به‌دست‌آمده و زمان کل، انتخاب می‌شود. نتیجه تعداد ۱۰۰ است. این فرایند در

شکل ۷ تا شکل ۹ نشان‌داده‌شده است. پارامتر ساختاری بعدی، تعداد حداقل نقاط است. مقدار حداقل تعداد نقاط موردنیاز برای تشکیل یک خوشه، تصادفی و به‌دلخواه ۱۰ انتخاب شده است. مقادیر پارامترهای روش پیشنهادی در جدول ۵ آمده است.

شکل ۷ نشان می‌دهد که از نظر خطای MAE، مقدار ۱۰۰ به کمترین مقدار دست‌یافته است.

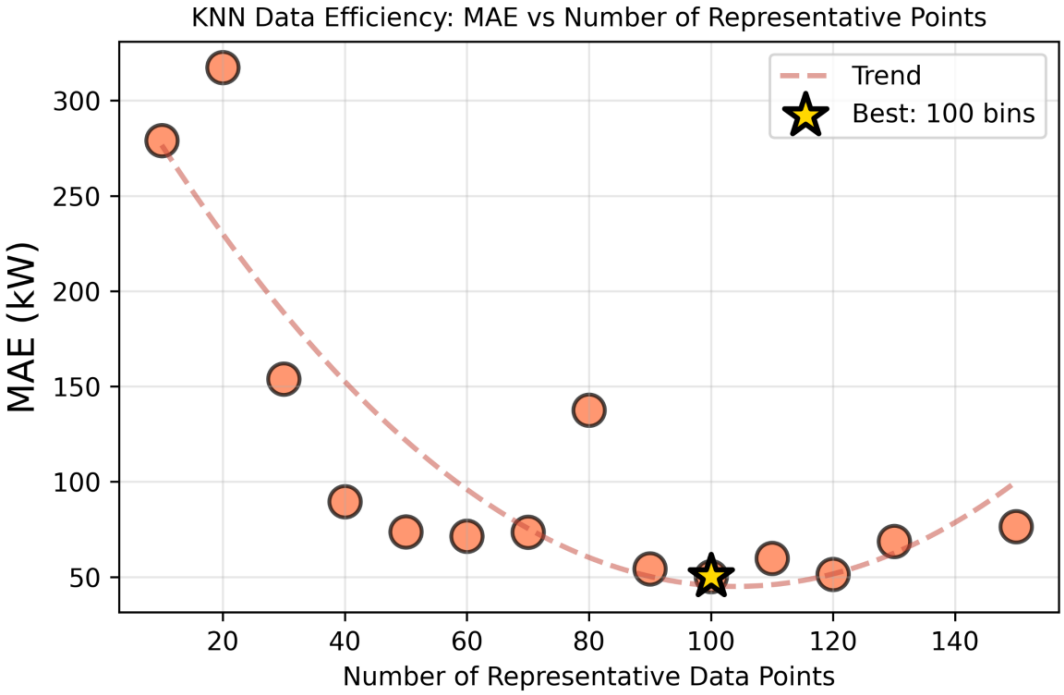
از طرفی، در شکل ۸، تعداد بازه‌ی ۱۰۰ نیز زمان آموزش متوسط را به نسبت دیگر مقادیر بازه‌ها به دست آورده است که باتوجه‌به سبک‌وسنگین کردن دقت مناسب آن، می‌توان از اختلاف زمانی اندک به وجود آمده چشم‌پوشی کرد.

به‌صورت کلی، شکل ۹ نمودار جامعی از عملکرد مدل‌سازی را بر اساس تعداد بازه‌ها نشان می‌دهد که همه‌ی شاخص‌های ارزیابی در روابط (۸) تا (۱۰) را در برمی‌گیرد. تعداد بازه ۱۰۰، در همه‌ی این شاخص‌ها به برتری



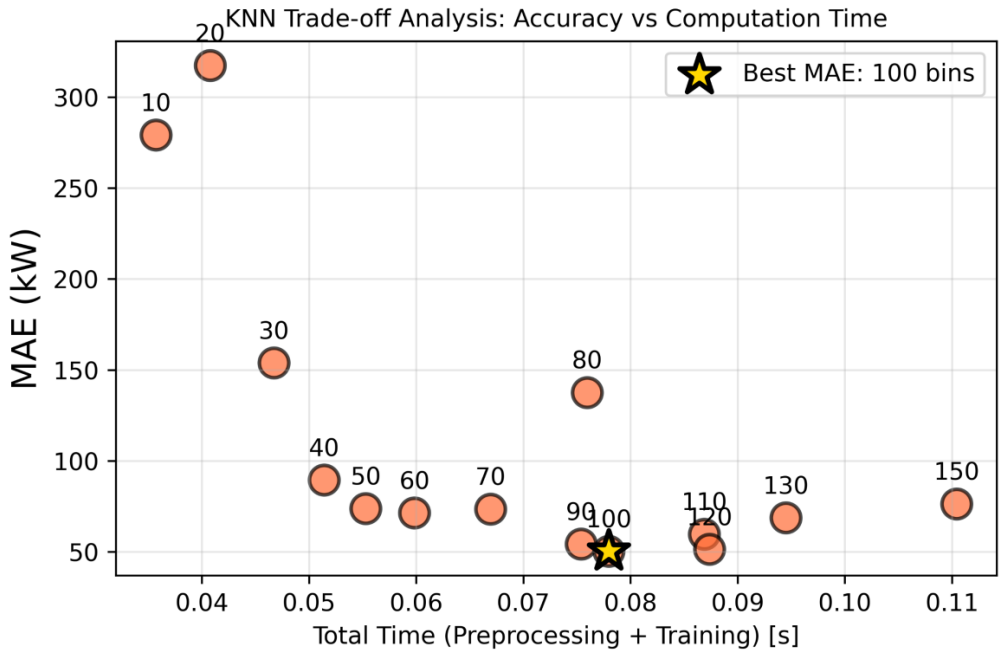
شکل ۶. منحنی توان پاک‌سازی شده توربین بادی با دی‌بی‌اسکن و استفاده از فیلتر.

Fig. 6. Cleaned wind turbine power curve using DBSCAN with filtering.



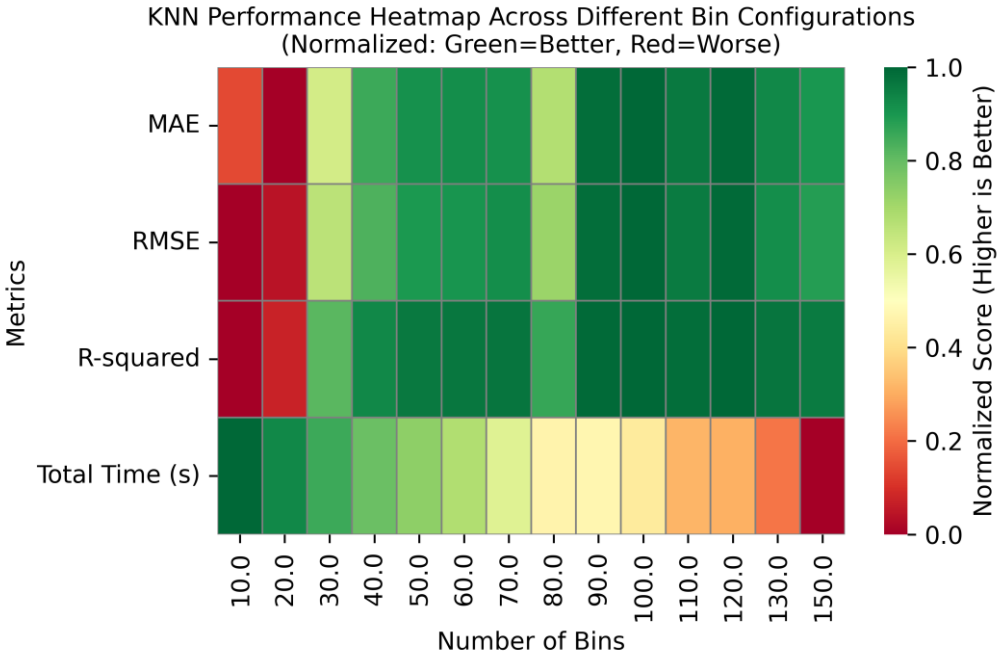
شکل ۷. نمودار تغییرات MAE بر اساس تعداد بازه برای انتخاب بهترین مقدار بازه.

Fig. 7. MAE variation based on the number of intervals for selecting the optimal bin value.



شکل ۸. نمودار تغییرات MAE و زمان بر اساس تعداد بازه برای انتخاب بهترین مقدار بازه.

Fig. 8. MAE and time variation based on the number of intervals for selecting the optimal bin value



شکل ۹. نمودار عملکرد مدل بر اساس بازه‌های متفاوت برای انتخاب بهترین بازه از نظر عملکردی.

Fig. 9. Model performance based on different bin number to identify the best number functionally.

جدول ۴. مشخصات ساختار دی‌بی‌اسکن و فیلتر.

Table 4. Structure of DBSCAN and filter.

همسایگی (ε)	۰/۵
حداقل تعداد نقاط موردنیاز برای تشکیل یک خوشه (min sample)	۵
شرط اول فیلتر	سرعت باد بین ۴ تا ۲۵ متر بر ثانیه و توان خروجی در محدوده‌ای بین ۰ تا ۱۰۰ کیلووات
شرط دوم فیلتر	سرعت باد بین ۱۴ تا ۲۵ متر بر ثانیه و توان خروجی بین ۳۳۰۰ و ۳۵۰۰ کیلووات

جدول ۵. مشخصات ساختار روش پیشنهادی جهت انتخاب نماینده.

Table 5. Structure of proposed method to select the representative data.

حداقل نقطه در بازه	۱۰
چندک (q)، ثابت برای هر توربین و مجموعه‌داده	۵۰
تعداد بازه‌ها (تعداد نقاط نماینده)	۱۰۰

رسیده است و مناسب‌ترین گزینه در بین تعداد بررسی شده است.

پس از به‌دست‌آمدن تعداد بازه‌ی مناسب (در اینجا ۱۰۰) داده‌های نماینده به نمایندگی از کل داده‌ها انتخاب می‌شود که در شکل ۱۰ نشان‌داده‌شده است. برای آموزش مدل به‌جای کل مجموعه‌داده، تنها از ۱۰۰ نقطه‌ی آبی‌رنگ استفاده می‌شود.

ب) آموزش مدل

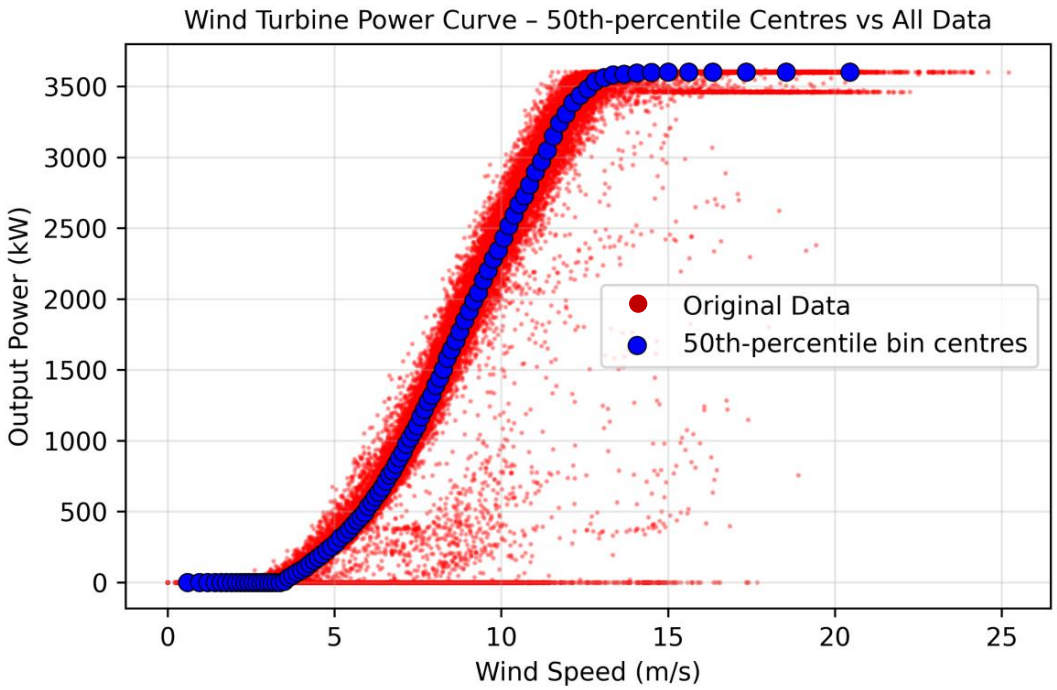
به‌منظور مقایسه‌ی بدون سوگیری روش‌های پیش‌پردازش، همه‌ی ساختار و عامل‌های دیگر نظیر ساختار مدل و مجموعه‌داده‌ی تست، به‌صورت مشترک و یکسان در نظر گرفته‌شده است و تنها تفاوت صرفاً روش‌های پاک‌سازی است. مدل K نزدیک‌ترین همسایه با داده‌های تمیز شده به روش‌های مختلف، آموزش می‌بیند و سپس همگی با یک مجموعه‌داده‌ی دیده‌نشده و مشترک بین دو روش، ارزیابی می‌شوند. ورودی مدل، سرعت باد انتخاب شده است و تعداد همسایگی عدد ۵ فرض می‌شود و فاصله از نوع اقلیدسی خواهد بود.

ج) ارزیابی عملکرد

برای ارزیابی کیفیت مدل‌ها، از معیارهای (۸) تا (۱۰) استفاده می‌شود. نتایج اولیه این ارزیابی و ناحیه‌ی اطمینان در شکل ۱۱ و شکل ۱۲ آمده است. طبق این شکل‌ها، مدل K نزدیک‌ترین همسایه به‌ازای هر دو روش پیش‌پردازش، عملکرد مناسبی از خود نشان داده است و اکثر تخمین‌ها در ناحیه‌ی اطمینان قرار داشته است؛ ولی تفاوت دقیق و پنهان وقتی ظاهر می‌شود که نه‌تنها نمودارهای بصری بلکه مقادیر کمی نیز مقایسه شوند. این مقادیر در جدول ۶ قابل‌مشاهده است. پاک‌سازی پیشنهادی تقریباً در همه‌ی موارد بهتر عمل کرده است و اوج عملکرد خود را در بهبود زمان‌های اجرا نشان می‌دهد به‌طوری که بیش از ۵۰٪ زمان کل کاهش‌یافته است.

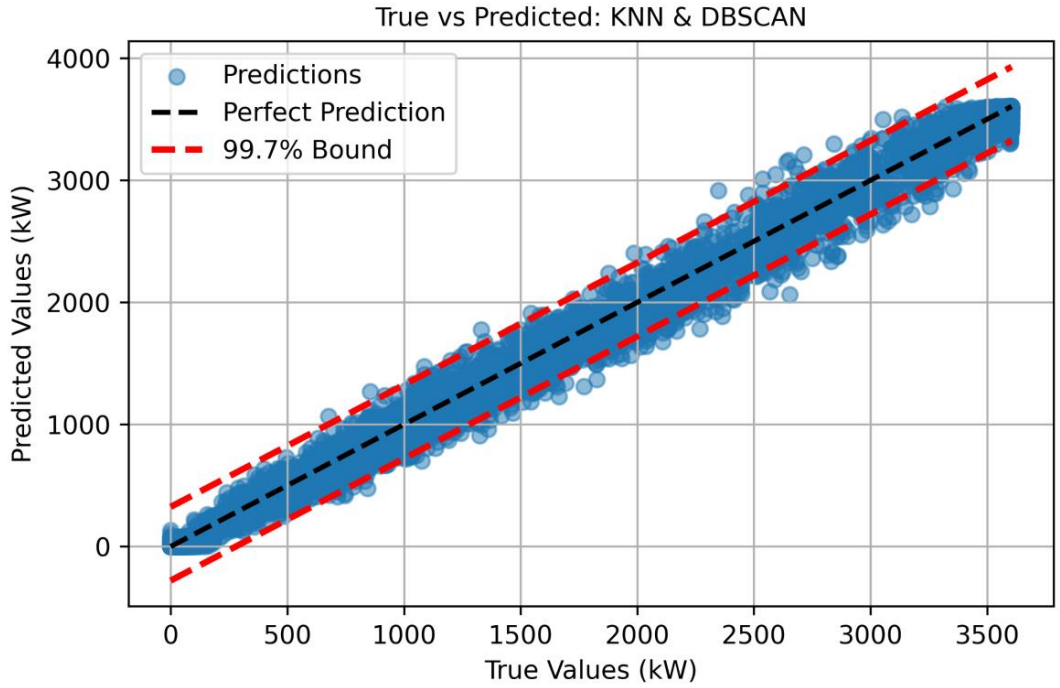
همچنین شکل ۱۳، کاهش زمانی را همزمان با نمونه‌ای از کاهش شاخص‌های ارزیابی (MAE) به‌صورت بصری نمایش می‌دهد.

در گام بعد تشخیص نابهنجاری، داده‌های تمیز شده به مدل داده می‌شود و مدل آموزش‌دیده، با استفاده از حدود آستانه‌ی چندکی محلی به تشخیص نابهنجاری می‌پردازد. برای نمونه، حدود آستانه‌ی محلی و سیگنال



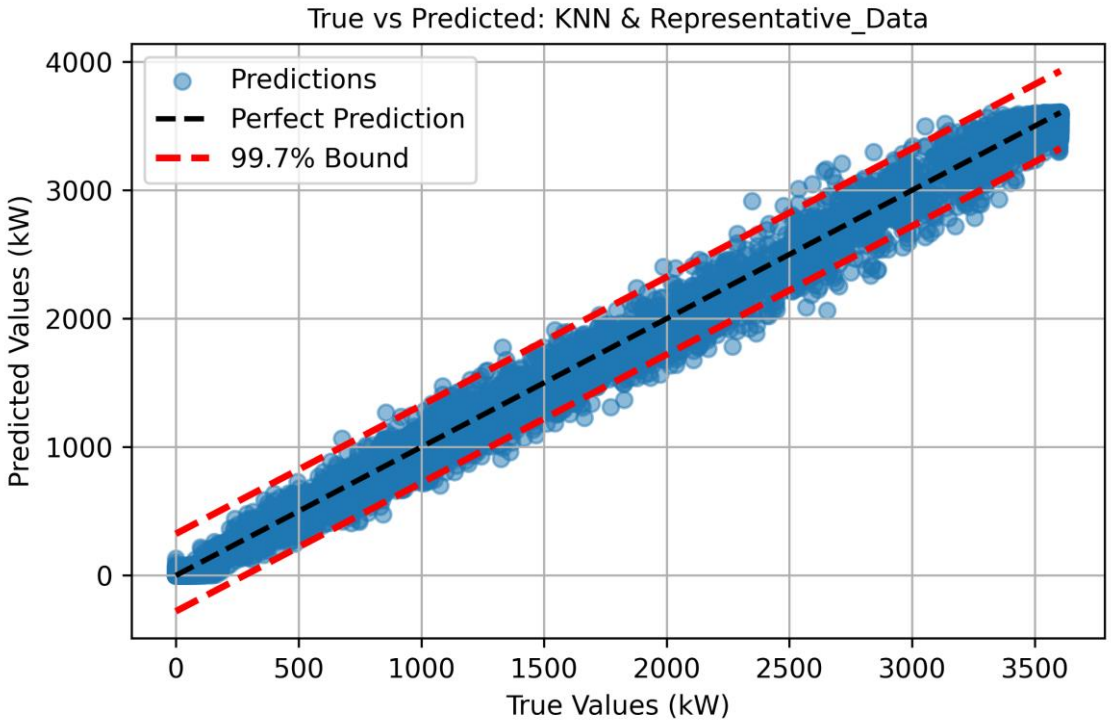
شکل ۱۰. منحنی توان و داده‌های نماینده.

Fig. 10. Power curve and the representative data.



شکل ۱۱. داده‌های پیش‌بینی شده و مقادیر واقعی برای دی‌بی‌اسکن و حد آستانه‌ی ۳-سیگما.

Fig. 11. Predicted data and true values for DBSCAN and 3-sigma threshold.



شکل ۱۲. داده‌های پیش‌بینی شده و مقادیر واقعی برای روش پیشنهادی و حد آستانه چندکی.

Fig. 12. Predicted data and true values for proposed method and quantile threshold.

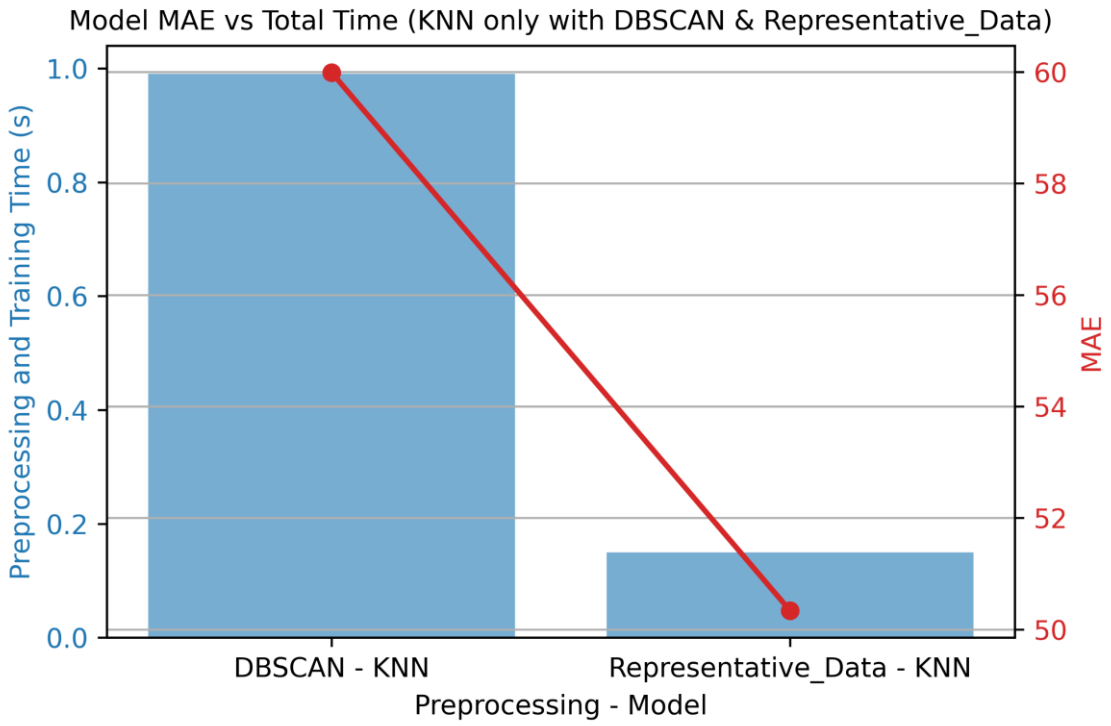
جدول ۶. نتایج و رتبه‌بندی مدل‌ها بر اساس عملکرد کلی.

Table 6. Results and ranking based on overall performance.

رتبه	مدل	پیش‌پردازش	RMSE	MAE	R^2	زمان مدل‌سازی (میلی‌ثانیه)	زمان پیش‌پردازش (میلی‌ثانیه)	زمان کل (ثانیه)
۱	K نزدیک‌ترین همسایه	روش پیشنهادی	۸۳/۰۵	۵۰/۳۳	۰/۹۹۵۸	۲	۱۴۸	**۰٫۱۵**
۲	K نزدیک‌ترین همسایه	دی‌بی‌اسکن	۹۸/۴۱	۵۹/۹۸	۰/۹۹۴۱	۱۰	۹۸۰	**۰/۹۹**

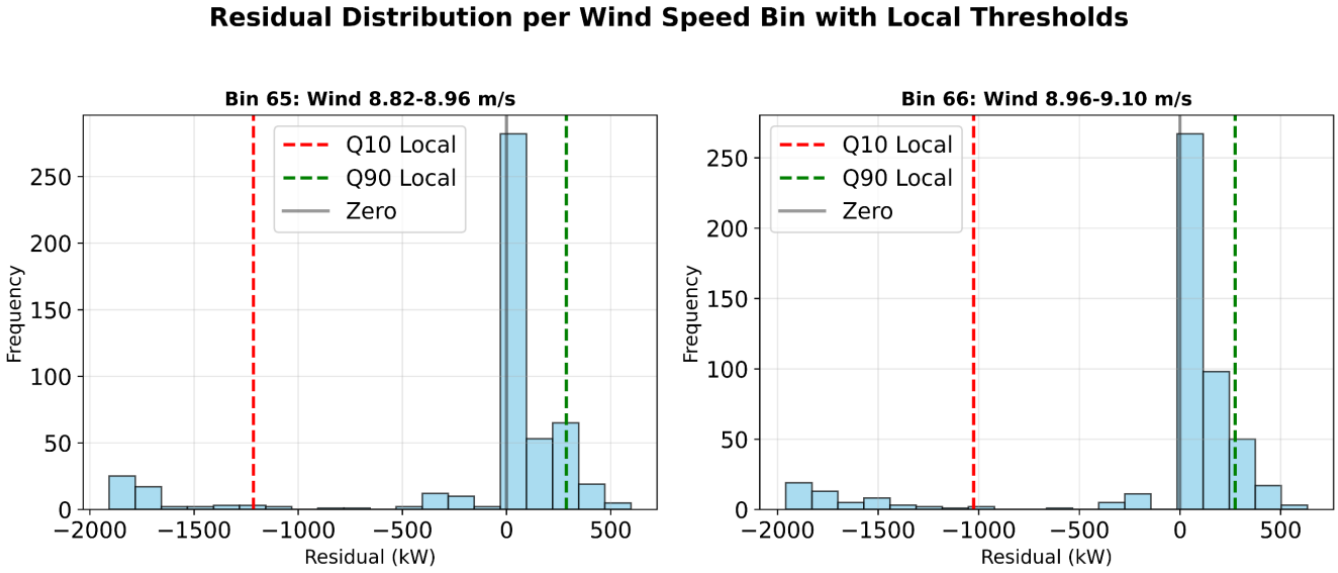
باقی‌مانده در بازه‌ی ۶۵ و بازه ۶۶ (سرعت باد حدود ۸/۸ تا ۹/۱) در شکل ۱۴ نشان‌داده‌شده است. این حدود، متناسب با هر بازه، متغیر و پویا است.

در شکل ۱۵ و نمودار سمت چپ، منحنی توان و داده‌های نابهنجار بر اساس حدود آستانه مشخص است. نمودار سمت راست، به حدود آستانه‌ی چندکی و سیگنال باقی‌مانده می‌پردازد. چنانچه داده تحت بررسی، خارج از بازه‌ی تعریف شده باشد نابهنجار و در غیر این صورت سالم اعلام می‌شود. این تشخیص و هشدار، نخستین لایه‌ی هشدار به متخصص بهره‌برداری جهت اتخاذ گام‌های نهایی است و در صورت سالم‌بودن داده، کارشناس از هشدار، صرف‌نظر خواهد کرد (به هر دلیلی مانند کاهش عمدی توان، کاهش بار در شبکه، کنترل حفاظتی، تعمیرات و... داده می‌تواند سالم باشد؛ ولی



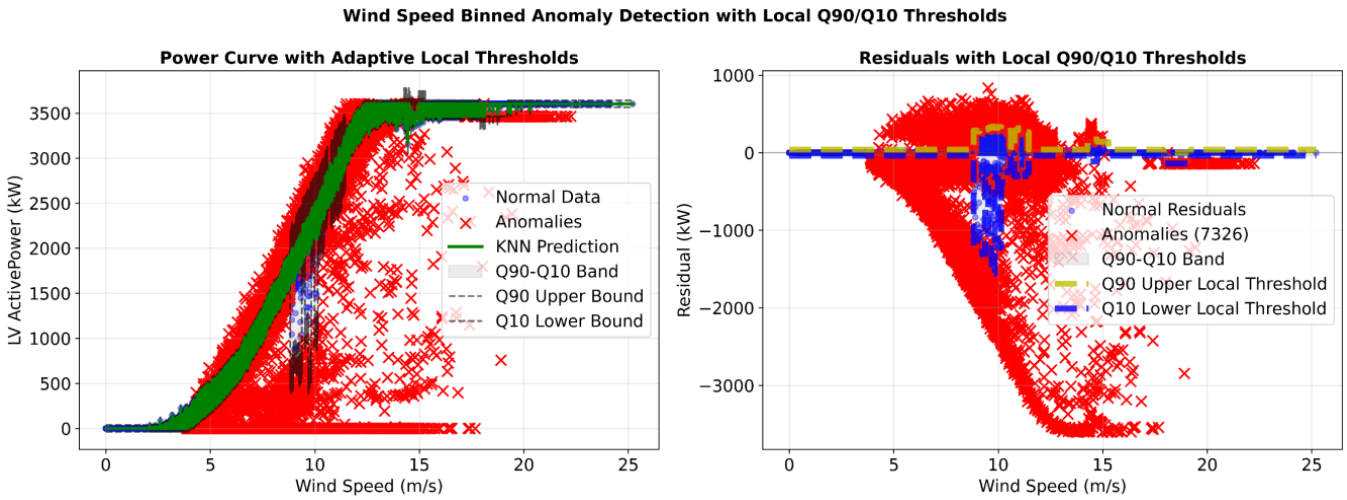
شکل ۱۳. نمودار کاهش زمان و MAE

Fig. 13. Time and MAE decrease chart.



شکل ۱۴. حدود آستانه چندکی محلی و سیگنال باقی در بازه‌ی ۶۵ و ۶۶.

Fig. 14. Local quantile thresholds and residual signal in 65 and 66 bins.



شکل ۱۵. مجموعه شکل‌های تشخیص نابهنجاری با استفاده از روش پیشنهادی و K نزدیک‌ترین همسایه به ترتیب از چپ به راست: نمودار منحنی توان و نمودار سیگنال باقی‌مانده-سرعت باد.

Fig. 15. Abnormal data detection for the proposed method and KNN algorithms. From left to right: the power curve and the residual–wind speed curve.

به‌صورت نابهنجاری ظاهر شود).

۵- نتیجه‌گیری

در این پژوهش، یک چارچوبی برای پاک‌سازی منحنی توان و شناسایی داده‌های نابهنجار در توربین‌های بادی ارائه شد.

در مرحله‌ی پیش‌پردازش، روش جدیدی معرفی شد. ابتدا، سرعت باد بازه‌بندی شده و میانه‌ی هر بازه به‌عنوان داده‌ی نماینده انتخاب می‌شود. این ساختار شامل تعداد بازه‌ها، چندک انتخابی و حداقل نقاط هر بازه است که چندک و حداقل نقاط می‌توانند ثابت در نظر گرفته شوند و تعداد بازه‌ها بر اساس استانداردهای موجود مانند IEC یا با آزمون‌وخطا تعیین گردد. این رویکرد ضمن کاهش حجم محاسباتی، دقت مدل‌سازی منحنی توان را حفظ می‌کند. در مرحله‌ی تشخیص نابهنجاری، به‌جای استفاده از آستانه‌ی سراسری ۳-سیگما، حدود آستانه‌ی چندکی محلی به کار گرفته شد. با بازه‌بندی سرعت باد و محاسبه‌ی آستانه‌ها بر اساس سیگنال باقی‌مانده، فرض نرمال بودن داده‌ها حذف شده و ناهمگنی خطا در سرعت‌های مختلف باد در نظر گرفته می‌شود.

نتایج آزمایش‌ها نشان داد که چارچوب پیش‌پردازش پیشنهادی، توانایی حذف مؤثر داده‌های پرت، کاهش زمان اجرا و ارائه‌ی منحنی توان دقیق‌تر را دارد و در مقایسه با روش دی‌بی‌اسکن، بدون نیاز به فیلتر اضافه، عملکرد بهتری در پاک‌سازی داده‌ها ارائه می‌دهد. به‌عنوان مسیرهای آتی پژوهش، می‌توان به ترکیب این چارچوب با الگوریتم‌های پیشرفته‌تر، استفاده از داده‌های محیطی و به‌کارگیری منحنی توان چند ورودی و رفتار نرمال اشاره نمود.

منابع

[1] C. Global Wind Energy, Global wind report 2023, Lisbon, 2023.

[2] C. Global Wind Energy, Global wind report 2024, Lisbon, 2024.

[3] F. Bilendo, H. Badihi, N. Lu, P. Cambron, B. Jiang, A normal behavior model based on power curve and stacked regressions for condition monitoring of wind turbines, IEEE Transactions on Instrumentation and Measurement, 71 (2022) 1–13.

[4] A. Llombart, J. Marques, J. Riera, Power curve characterization I: improving the bin method, in, International Conference on Renewable Energies and Power Quality (ICREPQ), 2005, pp. 367–371.

[5] R.K. Pandit, D. Infield, A. Kolios, Gaussian process power curve models incorporating wind turbine operational

model, Applied Energy, 296 (2021) 116913–116913.

[16] G. Liang, Y. Su, X. Wu, J. Ma, H. Long, Z. Song, Abnormal data cleaning for wind turbines by image segmentation based on active shape model and class uncertainty, Renewable Energy, 216 (2023) 118965.

[17] Z. Wang, L. Wang, C. Huang, A fast abnormal data cleaning algorithm for performance evaluation of wind turbine, IEEE Transactions on Instrumentation and Measurement, 70 (2021) 1–12.

[18] Y. Su, F. Chen, G. Liang, X. Wu, Y. Gan, Wind power curve data cleaning algorithm via image thresholding, in: 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2019, pp. 1198–1203.

[19] C. Zhang, D. Hu, T. Yang, Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost, Reliability Engineering and System Safety, 222 (2022) 108445–108445.

[20] G. Jiang, P. Xie, H. He, J. Yan, Wind turbine fault detection using a denoising autoencoder with temporal information, IEEE Transactions on Mechatronics, 23(1) (2018) 89–100.

[21] J. Wang, M. Kou, R. Li, Y. Qian, Z. Li, Ultra-short-term wind power forecasting jointly driven by anomaly detection, clustering and graph convolutional recurrent neural networks, Advanced engineering informatics, 65 (2025) 103137–103137.

[22] A. Meyer, Multi-target normal behaviour models for wind farm condition monitoring, Applied Energy, 300 (2021) 117342–117342.

[23] K. Leahy, R.L. Hu, I.C. Konstantakopoulos, C.J. Spanos, A.M. Agogino, D.T.J. O'Sullivan, Diagnosing and predicting wind turbine faults from scada data using support vector machines, International Journal of Prognostics and Health Management, 9(1) (2018) 1–11.

[24] A. Encalada-Davila, L. Moyon, C. Tutiven, B. Puruncajas, Y. Vidal, Early fault detection in the main bearing of wind turbines based on Gated Recurrent Unit (GRU) neural networks and SCADA data, IEEE/ASME Transactions on Mechatronics, 27(6) (2022) 5583 – 5593.

variables, Energy Reports, 6 (2020) 1658–1669.

[6] 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, International Journal of Energy and Environmental Engineering, 10(2) (2019) 181–188.

[7] W. Hernandez, A. Méndez, J.L. Maldonado-Correa, F. Balleteros, Modeling of a Robust Confidence Band for the Power Curve of a Wind Turbine, Sensors, 16(12) (2016) 1–13.

[8] M. Ester, H.P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226 –231.

[9] H. Yang, J. Tang, W. Shao, J. Yin, B. Liu, Wind power data cleaning using RANSAC-based polynomial and linear regression with adaptive threshold, Scientific Reports, 15(1) (2025) 5105.

[10] R. Morrison, X. Liu, Z. Lin, Anomaly detection in wind turbine SCADA data for power curve cleaning, Renewable Energy, 184 (2022) 473–486.

[11] Z. Mehmood, Z. Wang, Hybrid iForest-DBSCAN for anomaly detection and wind power curve modelling, Expert Systems with Applications, 289 (2025) 128381–128381.

[12] B. Jing, Z. Qian, H. Zareipour, Y. Pei, A. Wang, Wind turbine power curve modelling with logistic functions based on quantile regression, Applied Sciences, 11(7) (2021) 3048–3048.

[13] R. Koenker, G. Bassett Jr, Regression quantiles, Econometrica, 46(1) (1978) 33–50.

[14] C. Wan, J. Lin, J. Wang, Y. Song, Z.Y. Dong, Direct quantile regression for nonparametric probabilistic forecasting of wind power generation, IEEE Transactions on Power Systems, 32(4) (2017) 2767–2778.

[15] K. Xu, J. Yan, H. Zhang, H. Zhang, S. Han, Y. Liu, Quantile based probabilistic wind turbine power curve

[25] L. Xiang, X. Yang, A. Hu, H. Su, P. Wang, Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks, Applied Energy, 305 (2022) 117925–117925.

[26] S. Paudel, S. Faulstich, S. Sheng, Recommended key performance indicators for operational management of wind turbines, in, Journal of Physics: Conference Series, 2019, pp. 1–24.

[27] A. Aghajani Mobarakeh, A Review of Intelligent Fault Detection Methods for Wind Turbines, Iran University of Science and Technology (IUST), Tehran, Iran, 2024, (in Persian).

[28] E. Rashidian, Wind turbine fault diagnosis using SCADA data and normal behavior modeling, Iran University of Science and Technology (IUST), Tehran, Iran, 2023, (in Persian).

[29] T. Ouyang, A. Kusiak, Y. He, Modeling wind-turbine power curve: A data partitioning and mining approach, Renewable Energy, 102 (2017) 1–8.

چگونه به این مقاله ارجاع دهیم

A. R. Aghajani Mobarakeh, J. Poshtan, A Data-Driven Framework for Wind Turbine Power Curve Cleaning and Abnormal Data detection Based on Binning and Quantiles, Amirkabir J. Mech Eng., 57(10) (2026) 1263-1286.

DOI: 10.22060/mej.2026.24688.7894