Manuscript Number: wes-2025-131

Title: Failure classification of wind turbine operational conditions using hybrid machine learning

Dear Editor,

We have improved the article ref. wes-2025-131 was extensively revised in accordance with the reviewer's suggestions and comments. We strongly believe that Wind Energy Science would be the most suitable journal for this work. Please find attached the revised manuscript.

General comments from the reviewers. Reviewer 2 emphasised the relevance and motivation of the proposed study, while highlighting the strong potential for wind turbine monitoring. We are grateful for their constructive comments and suggestions, which have been carefully considered to improve the manuscript.

Our general response: This version has been carefully revised to incorporate all the reviewers' suggestions, and we hope it will be suitable for publication in the Wind Energy Science journal. In general, the reviewer's comments were highly appreciated and helped us significantly improve the quality of our manuscript. We are grateful for their consideration and time to review our paper. In the following pages, we provide point-by-point responses to each reviewer's comments. We have highlighted the revised parts in the manuscript in blue.

Yours sincerely,

M. R. Machado
A De Sousa
J. S. Coelho
R. Teloli

RC2: 'Comment on wes-2025-131', Anonymous Referee #2, 27 Aug 2025

The paper "Failure classification of wind turbine operational conditions using hybrid machine learning" By Machado et al. presents a hybrid machine learning framework for classifying wind turbine operational conditions, integrating supervised and unsupervised learning techniques. More specifically, the authors propose a novel relative change damage index for feature normalisation and apply canonical correlation analysis (CCA) for feature and sensor selection. The framework is implemented within the PyMLDA open-source platform and validated on experimental data from a small-scale Aventa AV-7 wind turbine. Results show excellent classification performance (up to 100% accuracy) in both binary and multiclass scenarios, with SVM outperforming other classifiers.

In summary, the proposed approach is relevant and well-motivated, and the results indicate strong potential for wind turbine monitoring. However, before being reconsidered for full acceptance, the following remarks should all be addressed by the Authors:

Answer: We thank the reviewer for the comments on our work and for the time to review our paper.

1. The title is a bit too generic and should be better detailed and circumscribed to the specific Machine Learning approach finally selected (Support Vector Machine).

Answer: The objective of this work is to propose a hybrid monitoring framework that combines multiple machine learning models and integrates multimodal data, thereby enhancing the interpretability of the Aventa wind turbine fault detection. Unlike conventional single-model or single-source approaches, this hybrid strategy leverages cross-domain correlations and environmental variability, enabling reliable monitoring under complex operational conditions. The model employs unsupervised k-means clustering to group data into homogeneous clusters, thereby facilitating pattern recognition without predefined labels, and multiple supervised classification machine learning algorithms for binary or multiclass fault classification. Since different algorithms may perform better under different scenarios and operating conditions, the proposed framework analyzes different machine learning algorithms. It identifies the best-performing model for the applied study case.

Therefore, restricting the title to a single specific method would not cover the main objective of this paper, which is better addressed in the revised manuscript. In this regard, the proposed title "Failure classification of wind turbine operational conditions using hybrid machine learning" is, in our view, more comprehensive and better reflects the nature of the study. Therefore, we believe the current title adequately represents the article's scope and objectives**.**

2. The study relies on a single small-scale 6.7 kW turbine dataset, partly with simulated faults (e.g., aerodynamic imbalance via roughness tape). The generalisability of the findings to utility-scale turbines in field conditions should be better acknowledged and discussed.

Answer: This study is associated with the ASCE-EMI Structural Health Monitoring for Wind Energy Challenge promoted by WedoWind and the Eastern Switzerland University of Applied Sciences. The primary goal of the challenge was to accurately detect three fault events: (1) pitch drive failure, (2) aerodynamic imbalance, and (3) icing. Among these, the aerodynamic imbalance was simulated using roughness tape, while the pitch drive failure and rotor icing were real faults that occurred in the turbine components. As part of this challenge, we have access only to the Aventa dataset and some information about the sensing position and period of the measurements. Although one of the failure conditions was induced, the others were genuine faults, ensuring realistic and diverse data characteristics.

The proposed data-driven model, developed and tested using a 6.7 kW turbine, effectively handles the intrinsic complexity of multi-source turbine data and fault variations representative of larger-scale systems. The methodology is designed to be scalable and adaptable, with a focus on its applicability to real-world operational scenarios. We acknowledge that the dataset originates from a small-scale turbine, however, the model was subsequently tested on a 1.65 MW turbine provided by an industrial partner, where it performed as expected. This reinforces the method's potential for generalization to utility-scale wind turbines operating under field conditions. This information was included in the conclusion section.

3. Related to the first remark, the robustness of the model under more complex environmental and operational variability (e.g., turbulence, large-scale icing, mixed faults) is not clear. More discussion on scaling the framework to real-world turbines is needed.

Answer: The current study considers general environmental and operational variability. Indeed, evaluating the influence of harsh environmental conditions, such as strong turbulence, large-scale icing, or mixed-fault scenarios, would be an important next step to validate the model further. Our group has prioritized a data-driven approach, relying on real measurements to develop and refine the methodology. However, we currently lack access to datasets capturing such extreme or combined operating conditions. We are actively seeking collaborations and data from turbines exposed to these environments to advance the study.

Nevertheless, the proposed framework was designed with adaptability in mind and can be extended to process data from real-world, utility-scale turbines operating under complex and dynamic conditions. Our ongoing studies focus on validating the approach using large-scale, more diverse datasets to assess further and enhance its reliability and generalization capabilities.

4. The claimed methodological novelties (relative change damage index and CCA-based feature/sensor selection) are interesting but not sufficiently differentiated from existing approaches in the literature. A stronger justification and comparative analysis would clarify their true contribution.

Answer: The methodological novelties of this study are embedded within the proposed hybrid monitoring framework, which integrates multiple machine learning models with multimodal data sources (structural and SCADA). This integration enables the framework to capture cross-domain correlations and operate reliably across varying environmental and operational conditions, advancing previous traditional single-source or single-model SHM approaches.

Within this framework, the relative change damage index introduces a feature normalising and scaling strategy that enhances the comparability of heterogeneous features without requiring predefined baselines. This improves sensitivity to operational deviations and ensures consistent feature interpretation across different sensors. Additionally, the canonical correlation-based feature and sensor selection method evaluates multivariate dependencies between response features and fault classes, providing a physically consistent, data-driven basis for ranking sensor importance. The main contributions of this study are:

(i) the development of a hybrid ML framework for operational fault assessment combining multiple algorithms and multimodal data,
(ii) the introduction of a feature relative change strategy for feature normalisation and scaling, and
(iii) the implementation of a canonical correlation-based feature and sensor selection process.

The proposed model enhances interpretability, scalability, and diagnostic performance. Comparative results across different scenarios confirmed the model's accuracy (85–98%) and stability, validating the methodological distinction in a practical application.

5. The reported near-perfect accuracies (often 100%) raise concerns of overfitting or data leakage. The authors should explain in more detail how independence between training and testing sets was ensured, and ideally provide confidence intervals or statistical robustness checks.

Answer*:* We thank the reviewer for the comment regarding the possibility of overfitting or data leakage, given the high accuracy reported. In the revised manuscript, further investigation was conducted, and all results were re-evaluated and validated.

The dataset is multiphysical, combining structural (accelerometers) and environmental (SCADA) information. It was found that when SCADA data were used directly in the k-means labelling stage, their strong internal correlation could bias group formation, leading to artificially high metrics. To eliminate this effect and ensure independence between the training and test sets, four different configurations of the dataset were evaluated, in which k-means labelling was applied to structural characteristics, with SCADA data used as the following variables, and the resulting

time-aligned data were subsequently used. In addition, all results were reprocessed using stratified cross-validation, and confidence intervals were calculated from multiple random divisions, ensuring statistical robustness. Thus, the high accuracy rates observed reflect the physical consistency and discriminative power of the selected features rather than overfitting or data leakage.

Included in the manuscript:

The final refined dataset organization is a crucial step in enabling machine learning algorithms to classify operational failures accurately. The deployed dataset exhibits a multimodal nature, integrating structural, temperature, and wind velocity information. The structural information is directly obtained from the accelerometers, which capture the system's physical response and potential failure signatures. In contrast, temperature and wind speed data are obtained from the SCADA system and represent indirect but relevant variables that influence the dynamic behaviour measured by the accelerometers, although they do not directly describe the damage dynamics. The final dataset is organized into four configurations, which are evaluated and their relevance demonstrated in the section 4.

- ***Fe-kms-Sc:*** *dataset composed of the selected features (Fe) extracted from the accelerometer time signals, followed by the SCADA data (Sc), and beelining provided by the k-means clustering (kms).*
- ***Fe-Sc-kms:*** *dataset including the selected features, the SCADA data, and the corresponding k-means labels.*
- ***Fe-kms:*** *dataset including only the selected features and the k-means labels, without incorporating the SCADA information.*
- ***Fe-kms-LoseSensor:*** *dataset including the selected features excluding those obtained from the most sensitive sensor identified by the Fast CCA method, together with the k-means labels and SCADA data. This configuration simulates the loss of the most sensitive sensor in each analysis scenario.*

The k-means labelling of the datasets Fe-kms-Sc, Fe-kms, and Fe-kms-LoseSensor is applied exclusively to the features that contain the most relevant physical information about the damage. The SCADA data are then organized by the corresponding day and hour for each feature sample. This ensures that the SCADA records follow the same row reordering imposed by the k-means clustering, while preserving the correct temporal (day and hour) correspondence between the SCADA measurements and the related accelerometer data.

The results for each dataset are shown in the new Figure 10, which compares the ML metrics across algorithms and datasets.

Included in the manuscript:

The ML models were tested on the Fe-kms-Sc, Fe-Sc-kms, Fe-kms, and Fe-kms-LoseSensor datasets to assess the influence of environmental dependencies (SCADA data) and the sensitivity of each sensor in failure evaluation. The metrics for the ***Fe-Sc-kms*** *dataset, shown as black (*) in Fig.10(a-d), indicate that, except for the NB classifier, all ML models achieved 100\% accuracy. Such perfect performance across multiple models can suggest potential issues like data leakage, overfitting, or improper dataset splitting. However, cross-validation with multiple random partitions was performed to ensure statistical robustness. In this case, the consistently high accuracy reflects the physical consistency and strong discriminative power of the selected features rather than methodological flaws.*

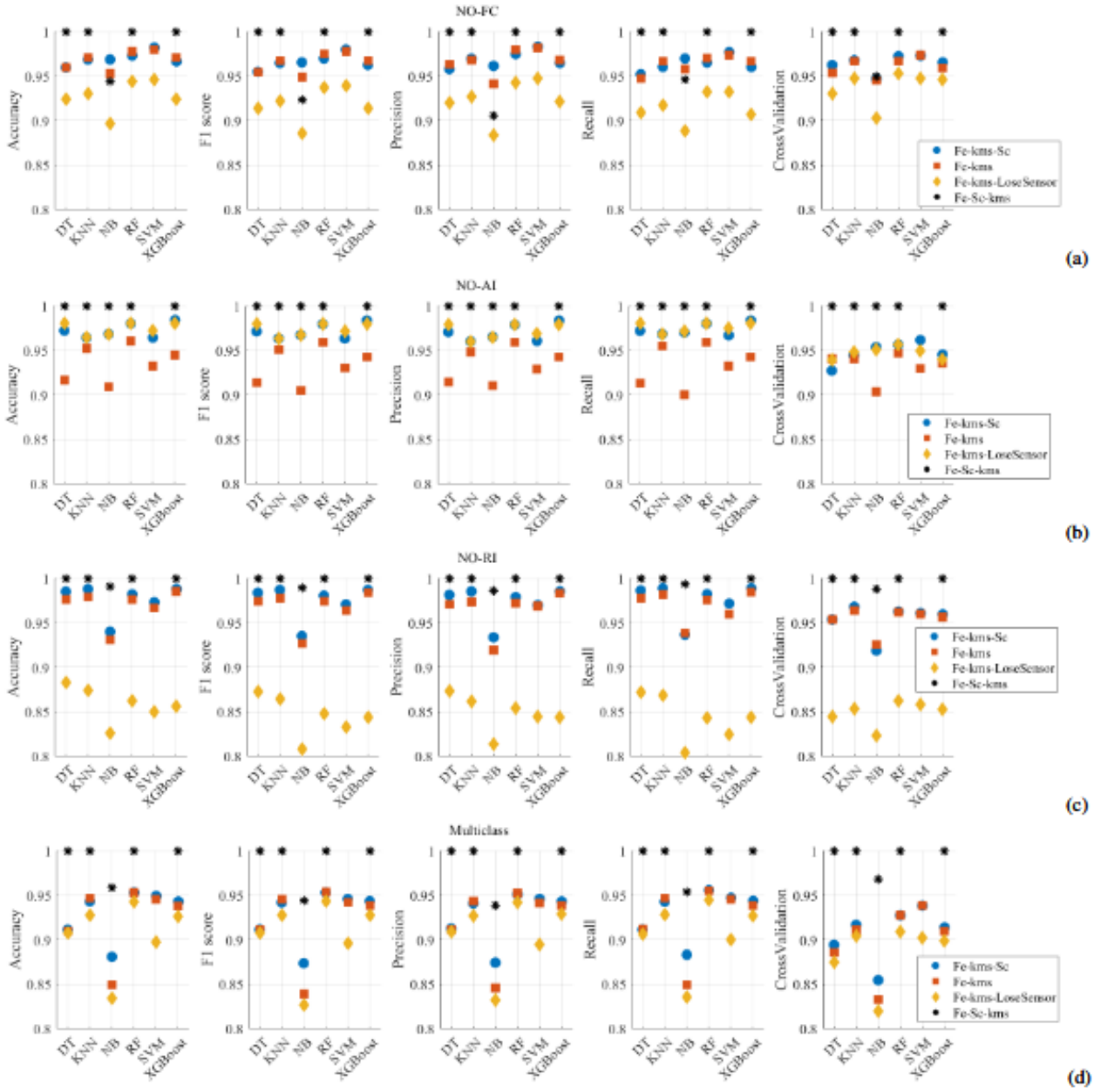

Figure 10. Comparison of the metrics (accuracy, F1-score, precision, recall, and cross-validation) for the six ML algorithms and each arranged final dataset. (a) Binary NO-FC, (b) Binary NO-AI, (c) Binary NO-RI, and (d) multiclass failure study cases

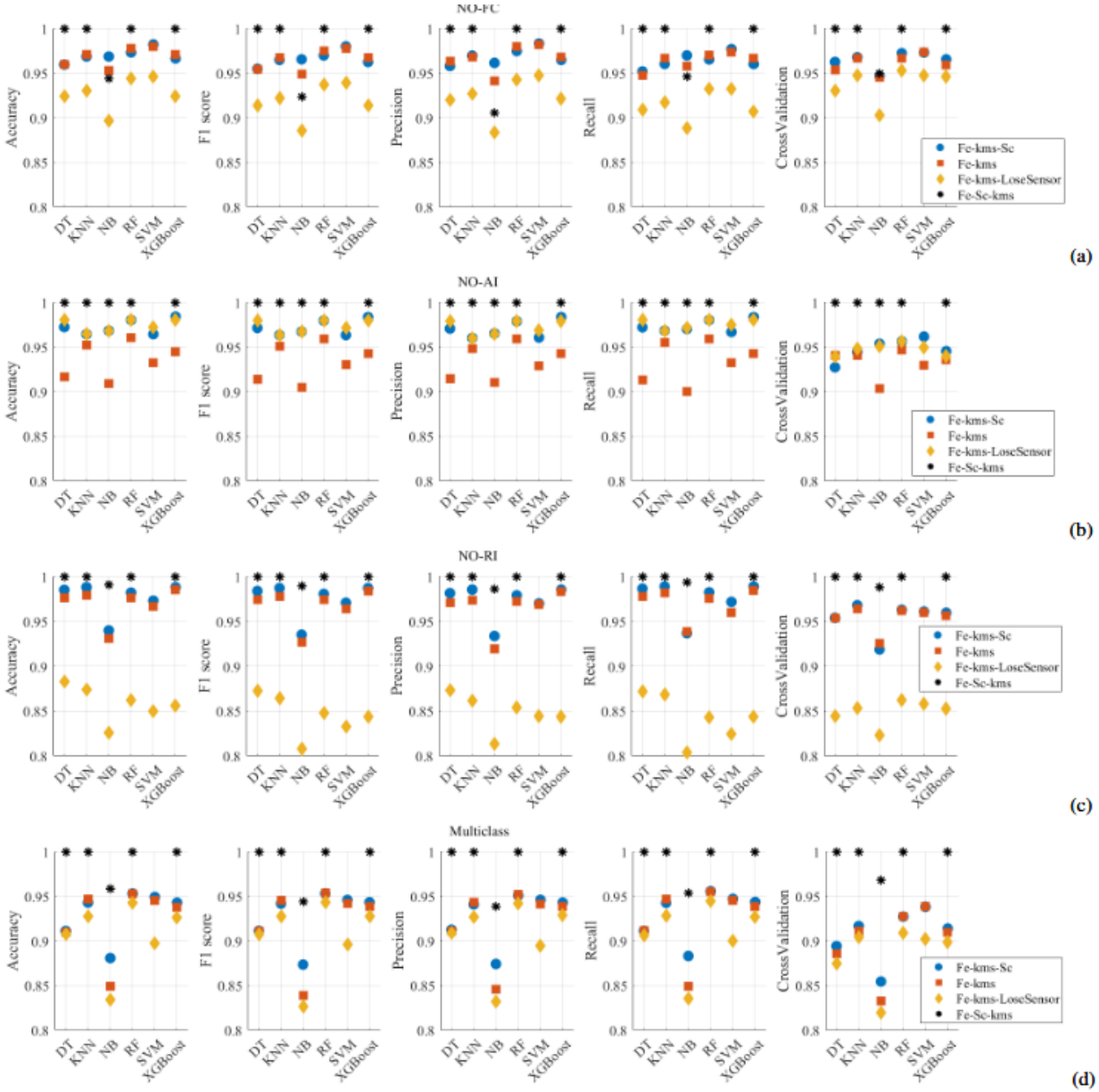

Figure 12. Sensor ranking score based on the top-performing features of each sensor for binary operational (a) NO-FC, (b) NO-RI, (c) NO-AI, and (d) for multi-classification.

The k-means clustering results and associated metrics also reveal clear class separation. The feature scores derived from the CCA, which quantify the linear association between the selected features and the k-means clusters (Fig.10), highlight the strong correlation between SCADA data and the structural sensors (accelerometers). The SCADA system provides environmental variables, such as daily temperature and wind speed, but lacks important structural information more directly related to damage states. In the Fe-Sc-kms dataset configuration, both features and SCADA inputs are used in the k-means clustering, which is predominantly influenced by the SCADA parameters. Therefore, the perfect ML metrics are attributed to dataset bias, where the models primarily classify failure conditions based on environmental variations rather than the structural response itself.

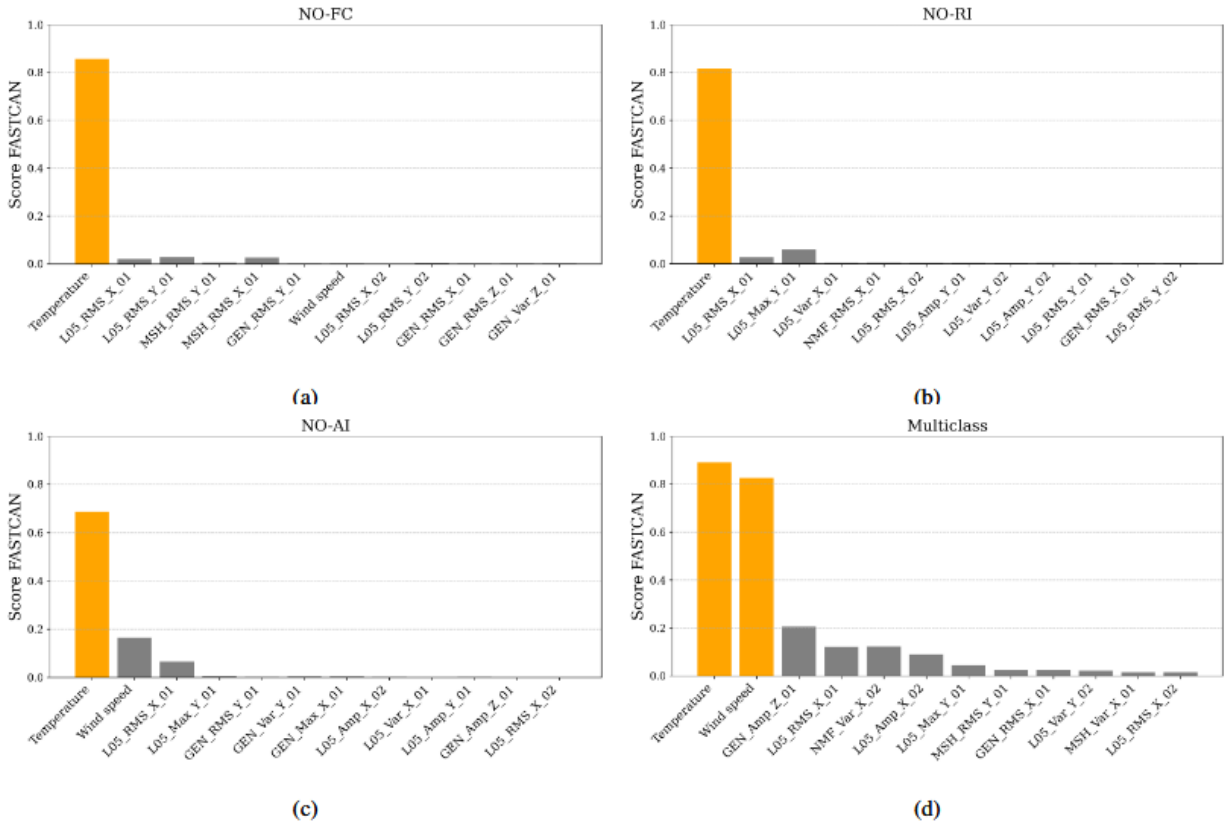

Figure 11. Feature score of the Fe-kms-Sc dataset for (a) NO-FC, (b) NO-RI, (c) NO-AI, and (d) multiclass.

*The ML models' metrics for the **Fe-kms-Sc** dataset, shown as blue $\bullet$ in Fig.11(a-d), indicate the sensitivity of the ML models to the damage. The k-means is performed on the structural features, and the SCADA follows its data reorganization, but it is not directly considered in the k-means labelling. For the binary case, the model's metrics range from 0.93 to 0.99. The SVM model reached 0.98 on the NO-FC case, XGBoost reached 0.98 for NO-AI, and XGBoost reached 0.99 for NO-RI. Multiclass case varied from 0.86 given by NB to 0.955 by RF.*

*The SCADA data are not included in the **Fe-kms** dataset, shown as orange-$\blacksquare$ in Fig.11(a-d). The performance metrics reached 0.98 with the SVM model for the NO-FC case, 0.955 with RF for the NO-AI case, 0.98 with XGBoost for the NO-RI case, and 0.95 with RF in the multiclass analysis. When the most sensitive sensor for each failure case was removed from the **Fe-kms-LoseSensor** dataset (yellow $\bLozenge $), performance metrics decreased, except for NO-AI, indicating the importance of this sensor's information for monitoring accuracy and model performance. This also reinforces that the most sensitive sensor is typically located near the damage site. In the NO-AI case, all sensors were positioned on the nacelle and tower. Although these sensors can capture the dynamic effects of blade aero-imbalance, they are distant from the local damage, which reduces their sensitivity. Consequently, the sensor group primarily captured global structural responses to the fault rather than local damage effects, explaining the performance variation when one of the sensors was removed.*

6. K-means clustering is used to initialise labelling, but the validation of clusters relies solely on the elbow method. Additional cluster quality indices (e.g., silhouette score, Davies–Bouldin) should be reported to strengthen confidence in the unsupervised stage.

Answer: We thank the reviewer for the suggestion regarding cluster validation. To ensure the quality of the unsupervised clustering with the K-means algorithm, we tested and included the suggested evaluation metrics, the silhouette score, and the Davies–Bouldin index.

For binary cases (k=2), the results indicate good separability between operating conditions, with consistent metrics across different components and failure types. In NO-FC scenarios, Silhouette Score values remain high (between 0.86 and 0.87) and Davies–Bouldin indices are low (between 0.12 and 0.13), indicating that points within each group are very similar to each other and that the clusters are well separated. In NO_RI and NO_AI conditions, the results still show a satisfactory clustering structure, with Silhouette values close to 0.56 to 0.60 and Davies–Bouldin between 0.58 and 0.75, indicating good cluster quality. In general, all binary-scenario cases yielded metrics superior to the reference values recommended in the literature (Silhouette > 0.5 and Davies–Bouldin < 1.0), confirming that the clustering method can adequately distinguish the different operational states of the turbines.
For the multi-class case with k=4, the validation metrics indicate good clustering quality. Silhouette Score values range from 0.52 to 0.60, and the Davies–Bouldin index values range from 0.50 to 0.60, indicating compact and well-separated clusters. These results confirm that the clustering process identified variations in turbine operational states. Therefore, these metrics provide strong evidence that the clustering step is reliable and that the labels assigned during the unsupervised phase are consistent.

Figure 5 was updated, including the K-means silhouette and its associated metrics. Figures 6, 7, and 9 follow the same organization but are for the other cases: NO-RI, NO-AI, and multiclass, respectively.

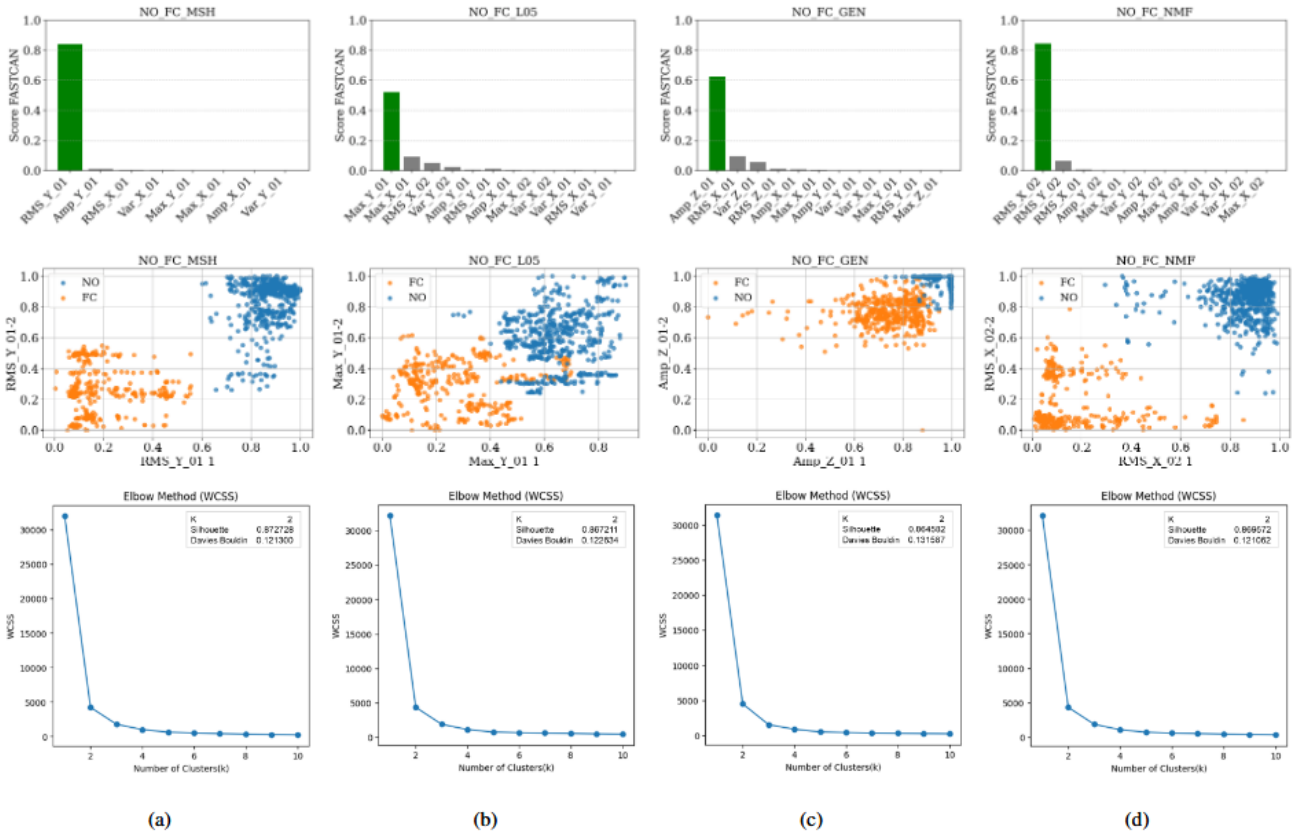

Figure 5. CCA feature score, dispersion diagrams of the highest-scoring feature, and k-means elbow method with respective metrics: (a) NO-FC sensor MSH, (b) NO-FC sensor L05 (1 and 2), (c) NO-FC sensor GEN, and (d) NO-FC sensor NMF (1 and 2).

7. The threshold of 0.6 for feature selection appears arbitrary; further justification or sensitivity analysis would be valuable.

Answer: The initial threshold value of 0.6 was set to ensure that the selected variables achieved at least 60% of the score metric. However, after further analysis, this threshold was removed in the revised version of the manuscript. Variable selection was then based solely on the highest canonical correlation scores. This criterion was applied consistently for both binary and multiclass cases.

8. The pipeline of Figure 1 is quite generic and, as it is now, does not really differentiate itself from any other ML-based Condition Monitoring approach.

Answer: Thank you for the comment. The pipeline in Figure 1 has been revised to clearly identify seven stages of the proposed monitoring framework, emphasizing the novelty associated with the hybrid multimodal model, feature and sensor selection, normalization, and multi-fault classification. Each step is now detailed in Sections 2.2, 2.3, and 3.2, highlighting how our approach differentiates from generic ML-based condition monitoring schemes. The final feature dataset and CCA (Fastcan) refinement process are described in Section 2.3, with results presented in Figs. 5–10.

Included in the manuscript:

*The monitoring framework consists of seven steps including clear identify in the revised pipeline figure (Fig.1) , where (1) receiving the acquired data**; (2) data processing and organisation**; **(3) feature extraction, normalisation, and grouping for similarity pattern**; (4) unsupervised feature labelling and clustering; **(5) feature and sensor selection**; **(6) data splitting, and ML failure identification and classification**; and **(7) Fault classification and model evaluation**. The final step also outputs the operational failure and identifies the best-performing ML algorithm based on its performance metric. The novelty associated with the hybrid model and multimodal data, feature and sensor selection, data normalization, and multiple fault classification is presented in a comprehensive set of steps outlining the process.*

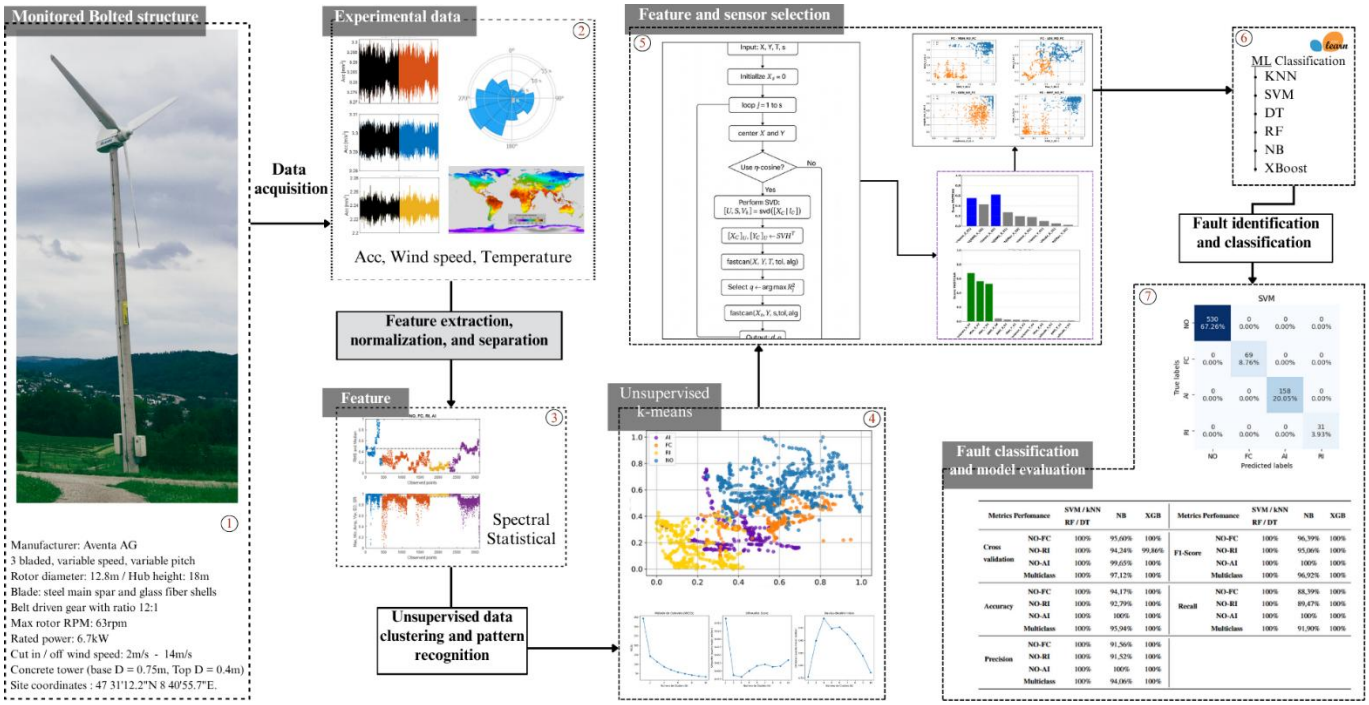

Figure 1. Pipeline of the hybrid machine learning model for fault classification on the Aventa 6.7 kW wind turbine.

The framework presentation and Fig.1 are followed by pseudo-algorithm 1.

9. The finding that SVM performs other classifiers fits well with recent findings in the SHM literature. However, Relevance Vector Machine (RVM) can be a feasible alternative, see e.g. https://doi.org/10.1016/j.oceaneng.2024.117692

Answer: We thank the reviewer for the suggestion regarding RVM. Indeed, SVM has demonstrated consistent performance in SHM-related studies, consistent with our results. We also recognise the relevance of RVM as a

promising alternative, as indicated in the suggested reference. In response, we have included a discussion of RVM in the vibration-based SHM section of the manuscript. The following description was added to the manuscript.

In addition to these approaches, recent studies have highlighted the potential of Relevance Vector Machines (RVM) as an alternative to SVM. RVM offers sparser solutions and greater computational efficiency without compromising accuracy, making it particularly suitable for scenarios with large data volumes and complex environmental variability (Kuai et al., 2024).

10. It would be useful to expand the context of Condition and Structural Health Monitoring for Wind Turbines, mentioning research review works, e.g. https://doi.org/10.3390/s22041627

Answer: We thank the reviewer for their suggestion. Indeed, expanding the context on Condition Monitoring (CM) and Structural Health (SHM) in wind turbines is pertinent. In response to the suggestion, we have included a brief discussion of review papers on the topic, with emphasis on https://doi.org/10.3390/s22041627. This addition reinforces the framing of our study within the current literature. The following description was added to the manuscript.

An effective wind turbine monitoring strategy combines Condition Monitoring (CM) of mechanical subsystems and SHM of structural elements (Civera and Surace, 2022). The authors highlight the growing integration of both into AI-based systems, aiming to improve reliability and operational safety.

11. While many figures are detailed, some are dense or partially redundant. Simplifying or condensing visual material, and clarifying captions to emphasise the main insights, would improve readability.

Answer: We thank the reviewer for their comment. In the revised version, we have simplified and condensed some figures, revised their captions, and highlighted the main insights to improve clarity and readability.

12. The conclusions somewhat overstate the generalisability of the proposed framework. A more balanced discussion of limitations (especially regarding dataset size, simulated faults, and applicability to large-scale turbines) would be appropriate.

Answer: We revised the conclusion section, incorporating the updated findings and the model's limitations as suggested.

Included in the manuscript:

This study proposes a data-driven hybrid framework for classifying operational conditions of a wind turbine, including normal operation, pitch-drive faults, rotor icing, and aerodynamic imbalance, encompassing multimodal data and multiple ML algorithms. The monitoring process follows an eight-step procedure: data processing, feature and sensor selection, feature normalisation, data splitting, unsupervised clustering, machine learning classification, and model evaluation. A novel relative change damage index was introduced to enhance scalability and normalise features extracted from structural dynamic responses and environmental conditions. Canonical correlation analysis was used to identify and rank the most sensitive features among the fifteen extracted from temporal responses and SCADA data (wind speed and temperature). Thus, multimodal information, including vibration signals from six accelerometers distributed across the turbine and environmental parameters (wind speed and temperature), was incorporated into the framework. Unsupervised k-means clustering enabled the discovery of homogeneous data groups, supporting robust pattern recognition without predefined labels.

Failure classification was implemented as both binary and multiclass tasks using kNN, SVM, DT, RF, Naive Bayes, and XGBoost. Models were tested on the Fe-kms-Sc, Fe-Sc-kms, Fe-kms, and Fe-kms-LoseSensor datasets to evaluate environmental influence and sensor sensitivity. Using the Fe-kms-Sc dataset, the models performed best, where

SVM achieved the highest accuracy for NO-FC (0.98), XGBoost for NO-RI (0.99) and NO-AI (0.98), and RF for multiclass classification (0.955). The Fe-kms-Sc dataset, which combines structural features with aligned SCADA data, yielded the most reliable failure detection. Excluding SCADA data (Fe-kms), the models focused on structural changes, with binary accuracies of 0.93-0.99 and multiclass accuracies of 0.86-0.95. Including SCADA alone (Fe-Sc-kms) produced perfect metrics due to environmental dominance, whereas removing the most sensitive sensor (Fe-kms-LoseSensor) reduced performance, confirming the importance of sensor placement near the damage and identification of the most sensitive sensor in each analysis.

The proposed hybrid model, developed and tested on a 6.7kW Aventa turbine, effectively manages the complexity of multi-source turbine data and representative fault variations. Its success relies on careful feature and sensor selection, ML model selection, inclusion of environmental data, dataset multimodality, and thoughtful dataset arrangement, which together enhance discriminative power and classification reliability. However, validation is limited by the small size of the wind turbine, the use of induced faults to simulate aero-imbalance in the blades, and differences between small-scale experimental turbines and large-scale operational turbines or wind farms. Extreme environmental and operational events, such as strong turbulence, large-scale icing, and mixed-fault scenarios, were not included in this study due to the controlled nature and limited scope of the experimental datasets. Ongoing studies aim to evaluate the methodology on larger, more diverse datasets, further explore environmental effects, extend the study to offshore wind farms, and assess reliability, generalisation, and applicability to utility-scale systems.