Referee #2

Comment 1

(1) The novelty of the predictive modeling component is not sufficiently clear. The manuscript mainly evaluates conventional machine learning models, with the best performance obtained from tree-based methods such as XGBoost and LightGBM. The authors should clarify why more advanced models for multivariate time-series analysis are not considered and better distinguish the methodological novelty from system-level integration.

Response 1

We sincerely thank the reviewer for this highly constructive and insightful comment. We agree that explaining the rationale behind excluding sequential multivariate time-series models and more clearly delineating our methodological novelties from system-level integration will significantly enhance the technical depth and clarity of our manuscript.

Time series analysis has not been used because of the necessity to temporally aggregate the exceptionally dense, high-frequency raw telemetry to overcome severe bandwidth and latency bottlenecks in our real-time IoT architecture. Far from compromising the analytical rigor, compressing this massive data successfully resolved the extreme sparsity caused by excessive zeros without any meaningful data loss, ultimately yielding a robust, balanced dataset perfectly optimized for our high-performance classification approach.

To address the reviewer's concern, we have incorporated these explanations into the revised manuscript:

- ***Section 3.2.1: Prototype Dataset and Descriptive Statistics (See Lines: 167-184)*** has been updated to detail the temporal aggregation of the dense raw telemetry and explain why sequential models were unsuitable.
- ***Section 3.5: Evaluated Models for Decision Support (See Lines: 245-255)*** has been revised to provide the mathematical and system-level justification for prioritizing tree-based ensembles over multivariate deep learning.
- ***Section 1.2: Problem Statement (See Lines: 48-65)*** and ***Section 2: Related Works (See Lines: 91-115)*** have been updated to clearly distinguish system-level contributions from methodological novelties.

Comment 2

(2) The role of the LLM and RAG components requires a much clearer explanation. It is unclear what knowledge base is used for RAG, how retrieval is performed, and what retrieved information is provided to the LLM. Based on the current description, the LLM appears mainly to serve as a natural-language interface for presenting sensor and prediction results rather than providing a new diagnostic or reasoning capability.

Response 2

We sincerely thank the reviewer for this valuable feedback. To clarify the role of the LLM and RAG components, we have added **Section 4 SWTFPS: LLM-Based Interactive User Interface *(See Lines: 355-365)***, which describes the knowledge base, semantic retrieval, real-time ThingSpeak data integration, and response generation processes. The RAG component retrieves relevant information from operator-provided wind turbine technical documents using Gemini embeddings and cosine similarity, while current sensor measurements and existing ML-based failure predictions are provided to the Gemini-based LLM. Importantly, the LLM does not perform failure prediction itself; instead, it interprets the existing prediction and current turbine conditions together with the retrieved technical knowledge. Thus, when an anomaly or failure is detected, the LLM goes beyond presenting monitoring results by generating a context-specific diagnostic recommendation based on the predicted condition, real-time sensor data, and relevant troubleshooting information.

Comment 3

(3) The manuscript lacks a clear unified framework or end-to-end workflow that connects the individual components. Although Figures 4 and 5 show the general system architecture and communication flow, the overall research methodology remains fragmented across SCADA preprocessing, model training, prototype sensing, pseudo-label generation, real-time inference, explainability, and the LLM interface. The manuscript would benefit from a unified methodological figure showing the complete workflow.

Response 3

We appreciate the reviewer's observation regarding the fragmented presentation of the methodology. To address this issue, we have added ***Figure 7: Integrated SWTFPS research pipeline*** under ***Section 3.5 Evaluated Models for Decision Support (See Lines: 251-255)*** as a unified end-to-end methodology workflow, which visually connects the major stages of the proposed framework and clarifies their relationships. The diagram illustrates how the methodological components are integrated into a single workflow, including the connection of the LLM interface to the real-time inference stage and the role of explainability in interpreting the model outputs.

Comment 4

(4) The literature review is relatively limited and does not sufficiently establish the research gap. Rather than mainly listing individual studies and their performance, the authors should provide a more systematic overview of SCADA-based prediction, multimodal sensing, IoT-based monitoring, explainable AI, and recent LLM/RAG applications in wind-turbine maintenance. A comparison table could also help distinguish the proposed work from existing studies.

Response 4

We thank the reviewer for the constructive feedback and have systematically restructured the literature review to cover the specified domains. Furthermore, we introduced a comprehensive comparison table ***(See Table 12 under Section 6.1: Interpretation of Results) (See Lines: 476-479)*** with a "Diagnosis through LLM Interface" column to explicitly establish the research gap and distinguish our proposed architecture.

Comment 5

(5) The claim that the newly added gas and sound sensors significantly improve fault prediction is not sufficiently supported by the current experiments. Feature importance and SHAP results indicate that the model uses these features, but they do not demonstrate that adding these sensors improves predictive performance. The authors should conduct an ablation study comparing model performance with and without the newly added sensors to directly quantify their contribution.

Response 5

We completely agree with the reviewer's methodological critique regarding the need to mathematically isolate the marginal improvements of the new sensors. To address this, we conducted a rigorous **Sensor Ablation Study** evaluating eight configurations, moving from a SCADA-equivalent baseline to models including localized sensors. Importantly, this study used an **independent validation dataset**, manually labeled using precise logs of physically injected failures (e.g., blade imbalances, friction) to ensure absolute ground-truth validity. The results, detailing the performance impact of each sensor, have been added to the revised manuscript as ***Section 5.3: Independent Empirical Validation and Sensor Ablation Analysis*** *(summarized in* ***Table 8: Ablation study evaluating the individual and joint contributions of localized IoT sensors against manually labelled, ground-truth-verified independent prototype data) (See Lines: 405-429)****.*

Table 11: Ablation study evaluating the individual and joint contributions of localized IoT sensors against manually labeled, ground-truth-verified independent prototype data.

Scenario	Sensor Configuration	Accuracy	Precision	Recall	F1-Score	OP Score
Scenario 1	Base Model	97.24%	97.07%	92.56%	94.76%	94.85%
Scenario 2	Base + Vibration	99.37%	99.53%	98.14%	98.83%	98.66%
Scenario 3	Base + Sound	98.49%	98.56%	95.81%	97.17%	97.20%
Scenario 4	Base + Gas	97.37%	97.55%	92.56%	94.99%	94.89%
Scenario 5	Base + Vibration + Sound	99.50%	100.00%	98.14%	99.06%	98.75%
Scenario 6	Base + Vibration + Gas	98.87%	98.13%	97.67%	97.90%	97.77%
Scenario 7	Base + Sound + Gas	99.12%	98.60%	98.14%	98.37%	98.26%
Scenario 8	Full-Model	99.62%	99.53%	99.07%	99.30%	99.09%

The ablation study demonstrates the specific value of each localized sensor, with vibration monitoring providing the most significant standalone boost (+4.07% F1-Score, +3.81% OP) by capturing structural oscillations indicative of mechanical wear. Acoustic monitoring acts as a complementary early-warning layer for high-frequency micro-friction, enhancing the baseline further (+2.41% F1-Score). While the gas sensor yields a marginal numerical improvement (+0.23% F1) due to its sparse activation only during rare catastrophic electrical hazards, its physical integration remains essential for comprehensive safety coverage. Ultimately, these modalities act in powerful synergy; the Full Model achieves peak performance across all metrics (99.62% Accuracy, 99.30% F1-Score, 99.09% OP, and a +6.51% boost in Recall to 99.07%), rigorously validating the physical and mathematical necessity of our multi-sensor data fusion approach.

We have incorporated these substantial additions and discussions into the revised manuscript:

- A new section ***Section 5.3: Independent Empirical Validation and Sensor Ablation Analysis (See Lines: 405-429)*** has been introduced to detail the experiment.
- ***Table 8: Ablation study evaluating the individual and joint contributions of localized IoT sensors against manually labelled, ground-truth-verified independent prototype data)*** has been added to summarize the quantitative metrics across all eight configurations.

Comment 6

(6) The current manuscript reads more like a technical project report than a research paper.

Response 6

We sincerely thank the reviewer for this valuable critique, which guided us in strengthening the manuscript from a descriptive technical report into a hypothesis-driven research paper. Following this recommendation, we have restructured the narrative to enhance its theoretical depth. The revised paper now presents a cohesive scientific framework and a higher level of academic rigor that aligns with scholarly standards.