Referee #1

First, we would like to clarify that your inability to access the revised manuscript previously was simply because the file had not yet been uploaded to the system while the discussion process was still ongoing. We sincerely thank you for taking the time to review our response document in the interim.

Furthermore, please note that due to the extensive incorporation of the new ablation study and the subsequent reformatting of the final manuscript, the specific line numbers provided in our earlier responses have inevitably shifted. However, we assure you that the structural integrity and the exact content of all prior revisions remain entirely intact and unchanged within the text.

Comment 1

I have not accessed the revised manuscript yet, but I have gone through the response document, and most of the points seem to have been addressed carefully. I nevertheless remain concerned about the rationale and evaluation of the proposed teacher–student procedure.

Very high student accuracy is not surprising when the student is trained using labels generated by the teacher, because the student is effectively learning to reproduce the teacher's decisions. The resulting error therefore mixes the teacher's labeling error on the prototype, the change from XGBoost to LightGBM, and the contribution of the additional sensors. It does not clearly demonstrate that vibration, gas, or acoustic measurements improve actual fault detection.

A more convincing design would reserve a subset of independently verified prototype labels completely unseen by both models and use this subset to evaluate both the transferred teacher and the student against the same true labels. If the student performs better, the improvement could provide evidence that the additional prototype-specific features contribute useful information; if it performs worse, this would suggest that they do not provide a reliable benefit.

Please clarify whether independently verified prototype labels exist for any portion of the data. If they do, why are they not used for independent evaluation? If they do not, then I would also like clarification on how the ground-truth labels used to train the original SCADA teacher were obtained, and how the authors establish that the teacher-generated prototype labels are sufficiently reliable to serve as targets for the student.

Response 1

We sincerely thank the reviewer for this profoundly insightful, constructive, and methodologically rigorous critique. We completely agree that evaluating the student model on teacher-generated pseudo-labels introduces a risk of "mimicry bias" and does not mathematically decouple the teacher's labeling errors from the actual predictive contribution of the newly introduced prototype-specific sensors.

To address this concern with the highest level of scientific rigor, we have executed the exact evaluation framework proposed by the reviewer. We are pleased to report that we have compiled a completely separate, independent prototype testing dataset. Unlike the primary training phase,

this validation set has been manually labeled to establish an absolute physical ground-truth benchmark.

This manual labeling was achieved by directly cross-referencing our precise physical experiment logs, which meticulously recorded the exact timestamps of every physical failure injection (such as blade imbalances, gear-meshing friction, and electrical short-circuits) on the prototype turbine.

Using this independent, ground-truth-verified dataset (which was completely unseen during the training phase of both models), we conducted a side-by-side comparative evaluation of the SCADA-equivalent Teacher (Scenario 1) and the fully integrated Student (Scenario 8). The complete quantitative results have been added to the revised manuscript as a new section: ***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)***.

Our findings provide direct, empirical answers to the reviewer's key questions:

1. Existence and Use of Independently Verified Ground-Truth Labels:

Yes, independently verified prototype labels do exist. They were derived from the physical logs of our controlled hardware experiments. In the revised manuscript, we have leveraged this independent dataset specifically as a validation benchmark to mathematically justify the teacher-student transfer and quantify the localized sensors' value, while preserving our main model's operational metrics.

2. Does the Student Outperform the Transferred Teacher on True Ground-Truth Labels?

1. Yes, significantly. As presented in the newly added ***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)*** of the revised manuscript:

- The Baseline SCADA-only Teacher Model (Scenario 1) evaluated against the true prototype labels yields an Accuracy of 97.24%, an F1-Score of 94.76%, and an OP Score of 94.85%, with a Recall of 92.56%.
- The Fully Integrated Student Model (Scenario 8) evaluated against the exact same true labels achieves an Accuracy of 99.62%, an F1-Score of 99.30%, and an OP Score of 99.09%, with a Recall of 99.07%.
- This represents a major +4.54% increase in F1-Score and a substantial +6.51% increase in Recall (rising from 92.56% to 99.07%). This quantitative improvement provides definitive empirical proof that:
- The student model does not merely "mimic" the teacher. If it did, its performance on true labels would be capped by the teacher's error rate (Scenario 1). Instead, the student successfully learns to correct the teacher's SCADA-limited misclassifications.
- The newly introduced vibration, acoustic, and gas sensors contribute vital, non-redundant physical information that is mathematically invisible to SCADA-equivalent thermodynamic variables, thereby significantly improving actual fault detection under true physical conditions.

3. SCADA Teacher Ground-Truth and Pseudo-Label Reliability:

The original SCADA teacher model (XGBoost) was trained on historical, ground-truth SCADA data from actual operating wind farms (spanning 89 cumulative years of operation across 36 turbines), where failure events were confirmed by field maintenance reports. The reliability of the teacher-generated pseudo-labels on the prototype was established via a high-confidence probability threshold filter (90%). The fact that the student model trained on these high-confidence pseudo-labels ultimately achieves a peak 99.62% Accuracy and 99.30% F1-Score when evaluated against manually labeled physical ground-truth validates the safety, transferability, and high diagnostic reliability of our teacher-student domain adaptation paradigm.

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

- ***Section 3.2.1 Prototype Dataset and Descriptive Statistics (See Lines: 167-184)*** has been updated to detail the compilation and manual ground-truth labeling of the independent validation dataset.

- A new section, ***Section 5.3: Independent Empirical Validation and Sensor Ablation Analysis (See Lines: 405-429)***., has been introduced, ***including Table 8: Ablation study evaluating the individual and joint contributions of localized IoT sensors against manually labelled, ground-truth-verified independent prototype data***

In addition, ***Figure 6: AI-supported wind energy fault prediction teacher-student transfer learning model*** under ***Section 3.5 Evaluated Models for Decision Support (See Lines: 245-249)*** has been added to clearly illustrate the proposed teacher–student transfer learning workflow and its role in AI-supported wind turbine fault prediction.