

REVIEWER'S REPORT

Manuscript No.: IJAR-59491

Title: Using Artificial Intelligence for the Early Diagnosis of Vehicle Malfunctions: A Leakage-Aware Evaluation and Meta-Analysis

Recommendation:

Accept as it is

Accept after minor revision.....

Accept after major revision ...✓.....

Do not accept (*Reasons below*)

Reviewer's ID: JPR- **098**

Rating	Excel.	Good	Fair	Poor
Originality		✓		
Techn. Quality		✓		
Clarity		✓		
Significance		✓		

Detailed Reviewer's Report

Decision: Accept after Major Revision

Major Comments

1. Clarification and validation of the blocked splitting strategy

The manuscript uses consecutive blocks of 500 records and a group five-fold split to simulate unseen operating conditions. However, the dataset is described as containing continuous acquisition runs for each class. The authors should explain precisely how the blocks and groups were constructed and assigned to the folds. Since the dataset is ordered by class and each class forms a continuous run, the current splitting strategy may not fully represent generalization to genuinely unseen operating conditions. A more rigorous temporal or run-independent validation strategy should be considered.

2. Investigation of potential data leakage in preprocessing

The manuscript identifies data leakage as a central motivation, but it does not sufficiently explain whether all preprocessing operations were fitted exclusively on the training folds. The authors should clarify how standardization, window generation, feature engineering, and any other transformations were performed. All data-dependent preprocessing must be

REVIEWER'S REPORT

fitted on training data only and then applied to the corresponding test data to prevent information leakage.

3. Explanation of the near-perfect binary fault detection results

The reported AUC values of 1.000 and 0.9994 under blocked validation are exceptionally high, particularly when four-class fault isolation performs considerably worse. The authors should investigate whether this performance is influenced by systematic differences between the normal and faulty acquisition runs, sensor operating ranges, or other class-specific characteristics. Additional analyses of feature distributions and false-positive and false-negative rates are necessary to establish whether the results reflect robust fault detection rather than dataset-specific separability.

4. Justification for merging fault types F2 and F3

The manuscript reports that merging F2 and F3 produces a three-class accuracy of 98.22% using all signals. However, the physical definitions of these fault types are not established in the public dataset, as acknowledged in the limitations. The authors should justify this merging strategy using domain knowledge or empirical evidence and explain the practical diagnostic meaning of the resulting classes. The conclusions should distinguish improved classification after label merging from demonstrated improvements in identifying the underlying physical faults.

5. Insufficient detail for experimental reproducibility

The manuscript lists six classifiers and some of their hyperparameters, but the complete experimental configuration is not provided. The authors should report the relevant hyperparameters, random seeds, software versions, fold assignments, and model-selection procedures. They should also clarify whether hyperparameters were selected using an independent validation procedure or adjusted after examining test performance. These details are necessary to reproduce the reported results.

REVIEWER'S REPORT**6. Statistical analysis of model performance**

The manuscript reports mean accuracy and standard deviation across five folds, but it does not provide sufficient statistical evidence to determine whether the differences between classifiers are meaningful. The authors should report fold-wise results and, where appropriate, use suitable paired statistical comparisons with effect sizes and uncertainty estimates. Because the folds may not be fully independent, the selected statistical tests and their assumptions should be justified.

7. Further justification of the early-diagnosis claim

The study uses steady-state fault conditions and explicitly acknowledges that incipient faults have not been evaluated. Consequently, the term “early diagnosis” may overstate what the experiments demonstrate. The authors should clearly distinguish early detection of an existing abnormal condition from prediction of a developing fault before it becomes detectable by conventional diagnostics. Claims regarding early warning should be limited to the experimental evidence, and validation on incipient-fault data should be identified as necessary future work.

8. Methodological concerns regarding the meta-analysis

The meta-analysis includes nine studies, with within-study variance approximated using a common effective sample size of 1,000 because test sizes were not consistently reported. The authors should provide a transparent study-selection procedure, explicit inclusion and exclusion criteria, and a table detailing each study's dataset, sample size, classification task, validation method, and reported performance measure. The use of a common effective sample size requires stronger justification, and the potential influence of this assumption on the pooled estimate and confidence intervals should be discussed in greater depth.

9. Comparability of effect measures in the meta-analysis

The manuscript combines reported accuracy values with an F1-score from one study and pools results from heterogeneous diagnostic tasks and data sources.

REVIEWER'S REPORT

Accuracy and F1-score are not interchangeable measures, particularly when class distributions and classification objectives differ. The authors should justify this conversion or perform sensitivity analyses excluding the study that contributes F1-score instead of accuracy. The implications of the very high heterogeneity ($I^2 = 98.47\%$) should also be emphasized when interpreting the pooled estimate of 96.50%.

10. External validity and practical deployment

The experiments use a single laboratory dataset collected from one engine, which limits generalization to different engine models, vehicles, driving conditions, and sensor configurations. The authors should strengthen the discussion of external validity and, if feasible, evaluate the proposed approach on an independent dataset. They should also discuss the practical implications of sensor availability, inference latency, false alarms, and the effect of real-world operating variability before suggesting deployment in vehicle diagnostic systems.

Minor Comments

These comments focus on clarity, presentation, reporting, and technical consistency.

1. Improve the abstract's clarity. The abstract contains several numerical findings but does not clearly distinguish the principal contribution of the experimental evaluation from that of the meta-analysis. The authors should summarize the objectives, validation protocols, principal findings, and limitations more concisely.

2. Strengthen the research-gap statement. The Introduction should explain more explicitly how the proposed evaluation differs from previous studies using the Engine Fault DB dataset. The novelty of combining leakage-aware validation, fault-detection analysis, reduced fault isolation, and meta-analysis should be stated clearly.

REVIEWER'S REPORT

3. Provide a comprehensive dataset description. The authors should include a table summarizing the 14 input variables, their units, their measurement sources, and their availability through standard OBD-II systems. The class distribution should also be presented clearly to facilitate interpretation of the experimental results.
4. Clarify the blocked validation procedure. The description of consecutive blocks of 500 records is insufficiently detailed for independent replication. The authors should specify how incomplete blocks were handled, whether blocks overlapped, and how the grouping was implemented in the five-fold validation.
5. Report additional classification metrics. Accuracy and macro-F1 are useful, but per-class precision, recall, and F1-score would provide a more complete evaluation, particularly for the four-class classification task. The authors should also report the false-positive rate and specificity for binary fault detection.
6. Explain the classifier configuration. The selection of 150 trees for random forest, ($k=5$) for k-NN, and the stated MLP architecture requires further explanation. The authors should indicate whether these settings were selected from previous studies, preliminary experiments, or systematic hyperparameter optimization.
7. Improve the presentation of figures. Figures 1–4 should use readable axis labels, consistent font sizes, clearly distinguishable validation protocols, and sufficiently large legends. The authors should ensure that all plotted values, error bars, and confidence intervals are legible in the final published version.
8. Expand the feature-importance analysis. The manuscript identifies concentration as the most important variable based on permutation importance. The authors should specify the exact evaluation metric used, explain how the permutation procedure was implemented, and discuss whether the importance ranking remains stable across different folds or validation conditions.

REVIEWER'S REPORT

9. Improve reference and terminology consistency. The authors should verify the completeness and formatting of all references, particularly publication details for recent studies. Terms such as fault detection, fault isolation, early diagnosis, random validation, and blocked validation should be defined consistently and used throughout the manuscript.

10. Strengthen the limitations and future-work section. The conclusion should more explicitly acknowledge the limitations associated with the single-engine laboratory dataset, unknown physical definitions of the fault classes, assumed meta-analysis variance, and heterogeneous study designs. Future directions should prioritize independent vehicle-level validation and evaluation on genuine incipient-fault data.