MULTIDISCIPLINARY SCIENTIFIC RESEARCH
BJMSR VOL 11 NO 2 (2026) P-ISSN 2687-850X E-ISSN 2687-8518
Available online at https://www.cribfb.com
Journal homepage: https://www.cribfb.com/journal/index.php/BJMSR
Published by CRIBFB, USA

EARLY FAULT SEVERITY PREDICTION IN DIGITAL-TWIN–ENABLED INDUSTRIAL SYSTEMS USING IMBALANCE-AWARE MACHINE LEARNING AND TEMPORAL FEATURE ENGINEERING

Oualid Ali [(a)1]

[(a)] *Head of Computer Sciences Department, College of Arts & Science, Applied Science University, Manama, Bahrain; E-mail: Oualid.ali@asu.edu.bh*

ARTICLE INFO

Article History:

Received: 6th October 2025
Reviewed & Revised: 6th October 2025 to 6th April 2026
Accepted: 8th April 2026
Published: 13th April 2026

Keywords:

Digital Twin, Predictive Maintenance, Fault Severity Classification, Imbalanced Learning, Temporal Feature Engineering

JEL Classification Codes:

O14, O32, L42

Peer-Review Model:

External peer review was done through double-blind method.

ABSTRACT

Continuous industrial telemetry in high dimensions generated by digital twin (DT) platforms offers opportunities for large-scale fault and predictive maintenance (PdM). However, the real-life applications of PdM remain very challenging. In this paper, we investigate the potential of early fault severity prediction using a five-year, hourly dataset of DT-derived data from January 2019 to January 2024. The data includes 38 variables, including measurements such as vibration, temperature, pressure, acoustic signals, operational load, maintenance and repair history, anomaly scores, fault probability estimates, and a multiclass fault diagnosis label of No Fault or Critical Fault. The research presented in this paper follows an end-to-end analytical approach, using a set of machine-learning classifiers, then optimizing them using class-weighted classifiers, probability calibration, and a cost-sensitive decision policy, all aligned with the strategic goals of PdM. The results of the experiments conducted in this research show that the baseline and temporally enhanced models achieve ≈0.69 accuracy. When class-weighted learning is introduced, it increases fault recall to 0.059 and macro-F1 to 0.451, while threshold optimization achieves perfect fault recall (1.00) with a corresponding fault F1-score of 0.468. The main findings of this work show that the joint effects of temporal modelling, imbalance-aware learning, and calibrated decision thresholds govern early severity prediction performance. Our research is expected to yield a reproducible, deployment-oriented framework that improves minority-class detection and enables effective maintenance decisions for DT-enabling Industry 4.0/5.0 applications.

INTRODUCTION

The rapid adoption of Industrial Internet of Things (IIoT) infrastructure and Digital Twin (DT) technologies has turned modern industrial systems into data-centric ecosystems that can continuously monitor and support intelligent decision-making. DT-based predictive maintenance (PdM) leverages synchronised virtual images of physical equipment to analyse multivariate sensor feeds, operating parameters, and environmental factors. It aims to predict failures before growth into expensive or risky incidents (Abd Wahab et al., 2024; Grieves & Vickers, 2017). Compared to the conventional methods of reactive or schedule-based maintenance processes, PdM has proved to minimize the occurrence of unwanted downtime, enhance asset utilization, and enhance system safety in a wide range of industrial environments (Compare et al., 2020; Carvalho et al., 2019). Despite these steps, reliable fault prediction at early stages remains a daunting task. Empirical industrial data exhibit strong temporal correlations, nonlinear relationships among variables, and very skewed distributions of classes, with normal operating conditions predominating and fault incidents, especially severe or critical faults, being rather infrequent (Dalzochio et al., 2020). As a result, predictive control leads many machine-learning models to achieve spuriously high accuracy on the majority of values for a single class (typically no-fault). It thus does not provide meaningful early warnings of impending failures. Previous literature has emphasized that traditional evaluation metrics, such as accuracy or ROC-AUC, may be inaccurate in such situations. Metrics based on precision, recall, and cost-effective decision-making are more suitable for PdM use (Saito & Rehmsmeier, 2015; Sofaer et al., 2019).

Mathematically, where X t is the vector of sensor measurements, operational variables and other variables of operational condition and environmental forces, represented in the form of the DT derivative, i.e. Xt Є Rd, assume that yt ={ 0 1 2 3 4 }sufferings of the fault, i.e. no fault, critical fault. The key issue that was presented in this paper is to learn a predictive variable that predicts the historical data (Xt, Xt-1,…, Xt-k) to an accurate forecast of the future fault state yt + Δ

[1]Corresponding author: *ORCID ID: 0009-0001-1433-017X*

https://doi.org/10.46281/bjmsr.v11i2.2851

To cite this article: Ali, O. (2026). EARLY FAULT SEVERITY PREDICTION IN DIGITAL-TWIN–ENABLED INDUSTRIAL SYSTEMS USING IMBALANCE-AWARE MACHINE LEARNING AND TEMPORAL FEATURE ENGINEERING. *Bangladesh Journal of Multidisciplinary Scientific Research*, *11*(2), 28-36. https://doi.org/10.46281/bjmsr.v11i2.2851

, where Δ k is a prediction horizon. The formulation is specifically aimed at predicting early fault severity rather than fault diagnosis at the point of occurrence, and is thereby a more difficult but operationally valuable task (Compare et al., 2020; Zonta et al., 2020).

The literature on PdM and DT analytics has so far focused on the so-called remaining-useful-life (RUL) estimation, binary failure detection, and fault diagnosis using instantaneous features and randomly shuffled training and test partitions (Carvalho et al., 2019; Dalzochio et al., 2020; Li et al., 2018). Although high classification scores are often reported using such methods, their test conditions can unintentionally introduce temporal leakage or majority-class bias, thereby failing to reflect realistic deployment conditions. In the recent methodological reviews, it has been highlighted that time-conscious validation, temporal feature engineering and imbalance-conscious learning are needed to provide realistic and decision-relevant performance estimates (Abd Wahab et al., 2024; Saito & Rehmsmeier, 2015; Richardson et al., 2024).

This paper is motivated by these gaps and adopts a research design that aligns with established PdM machine-learning pipelines while emphasizing the importance of time and deployment relevance. Based on a five-year, hourly, DT-based industrial data set of 38 features across the sensing, operational, environmental, and maintenance dimensions, a variety of classification models are benchmarked and systematically refined with respect to temporal feature construction, class-weighted learning, and decision threshold calibration. The analysis aims to improve the ion-class detection and early-warning function for improvability, rather than maximizing accuracy, thereby making the model's viability acceptable for comfortable alignment with realistic maintenance goals.

The results of this work have direct implications for maintenance engineers, reliability analysts, and industrial decision-makers seeking reliable early-warning systems, as well as for developers of DT and IIoT solutions that aim to incorporate machine-learning models into operational monitoring systems. This research paper presents a comprehensive and valid approach to early fault severity prediction in an industrial setting, explicitly addressing class imbalance, temporal dependency, and evaluation bias to provide a reproducible and realistic framework. Based on this, the research question that will be the focus of this study is:

RQ: To what extent can imbalance-conscious machine-learning predictors, using temporal features engineering, make future fault severity predictions in Digital-Twin-based industrial systems in realistic time-aware evaluation conditions?

LITERATURE REVIEW

Recent predictive maintenance (PdM) studies have shifted the paradigm from detection when failed (reactive) to an anticipatory one, driven by the spread of Industry 4.0 sensorization and the growing popularity of digital twins (DTs). The surveys dedicated to DT-enabled PdM always claim that the main value of DTs is based on three facts: (i) higher-fidelity context, making it possible to mix physics-based models with operational data; (ii) the possibility to make a simulation of rare fault cases controllable; and (iii) the creation of closer feedback loops between virtual and real assets. However, these publications also bring up the enduring problems of data asymmetry, the gap between the virtual and the real world, and unreliability of assessment in case of missing the time leakage (Zhong et al., 2023; Abd Wahab et al., 2024; You et al., 2022; Semeraro et al., 2021).

Surveys and models related to DT-to-PdM have highlighted the significance of integration architectures, which are usually captured in the sequence IoT -data fusion -simulation/analytics -decision layer- and repeatedly mention that reports of high accuracy are most likely to emerge when the architecture is tested under simplified conditions, e.g., small scale studies, limited fault modes or non-deployable data splits (Zhong et al., 2023; Abd Wahab et al., 2024; You et al., 2022; Semeraro et al., 2021). As a matter of fact, normal operation prevails in industrial PdM, and fault classes (especially early-stage or infrequent severe faults) are very unobservable. The DTs can provide a solution by constructing artificial fault traces, but this can only be done when the artificially generated samples are physically plausible and conform to the empirical distributions of measurements (Long et al., 2026; Zayed et al., 2023; Yang et al., 2023).

Classical machine-learned baselines such as random forests, extra trees, XGBoost, and linear models also still hold up well in cases where features are well-crafted. Nevertheless, temporal degradation pattern modeling is required in many industrial PdM tasks where deep sequence models (CNNs, TCNs, LSTMs, transformers, and autoencoders) often perform better than feature-only pipelines, particularly in early warning and multivariate sensor fusion (Serradilla et al., 2022; Jaenal et al., 2024; Hancock et al., 2023; Cook & Ramadas, 2020). Surveys have, however, found that the most appropriate model is greatly dependent on the factors including windowing strategy, definition of label (fault-at-time t and within horizon), and the use of drift and imbalance (Serradilla et al., 2022; Jaenal et al., 2024). Other work in security and other classification challenges also investigate such tasks.

Accuracy is a poor performance measure in datasets with extreme class imbalance. The metrics like PR-AUC and class-wise F1 (or macro-F1) are more informative in skewed conditions (Zhao et al., 2022). One of the most common traps is setting models to optimize majority recall while sacrificing minority recall, i.e., tuning them to maximize majority recall. Recent work therefore advises the combination of cost-sensitive learning, threshold calibration specific to early detection and resampling algorithms specific to time-series (Hancock et al., 2024; Lyu et al., 2021; Isbilen et al., 2025).

Evaluation leakage is one of the main factors that lead some publications to achieve high F1 scores. Time-dependent splits may randomly cut off data time indices, potentially allowing future regimes, maintenance activity, or repeating machine conditions to enter the training set and thereby inflating performance measures. Empirical studies have shown that adequate time-based splits often reduce apparent scores but provide a clearer indication of deployment performance (Tao et al., 2019). In the case of PdM, this problem is further exacerbated when one uses maintenance logs, records of previous faults, or health index variables that might partially encode the label if they are not properly lagged (Tao et al., 2019). The methodological foundations of decision-tree-based predictive maintenance in the presence of class

imbalance are increasingly being clarified in modern research. Pu and Wu (2024) and Apeiranthitis et al. (2024) reported strong results in fault diagnosis using decision trees for reducers and rotating equipment. Although they focus on the post-manifestation aspect, this makes the problem less challenging than warning predictions. Imbalance-oriented approaches, such as SMOTified-GAN, very similar to the SMOT-inspired anchor construction scheme (Alobaid & Corcho, 2024), have demonstrated enhancement in minute class recall; however, in follow-up approaches, caution was raised that oversampling can cause the corruption of temporal relations among temporal time series data (Lyu et al., 2021). Studies on the suitability of metrics show that conventional accuracy and AUROC can be misleading under skewed distributions, prompting a shift towards PR-AUC and class-wise examination (Karyofyllas et al., 2025). Recent decision-tree-based deep learning models (Tan et al., 2025; Yoo, 2025; Mateus et al., 2025) have improved robustness but remain susceptible to choices of window length and label horizon, highlighting the need for careful temporal validation, as in this manuscript.

MATERIALS AND METHODS

In this section, we present the methodology used in our research. Figure 1 shows the main steps of the proposed approach.

The Dataset and Problem Formulation

The existing study uses a Digital Twin-based industrial dataset comprising 5 years of hourly observations from January 2019 to January 2024 (*IndFD-PM-DT*, 2024). This corpus contains thirty-eight variables, which include machine condition (e.g., vibration, temperature, pressure, acoustic signals), operational behaviour (e.g., motor speed, torque, load, utilization), environmental conditions (e.g., ambient temperature, humidity, air quality) and maintenance-related variables (e.g., previous faults, repairs, downtime). The main target variable, Fault Diagnosis, is a multiclass variable representing fault severity levels from No Fault to Critical Fault.

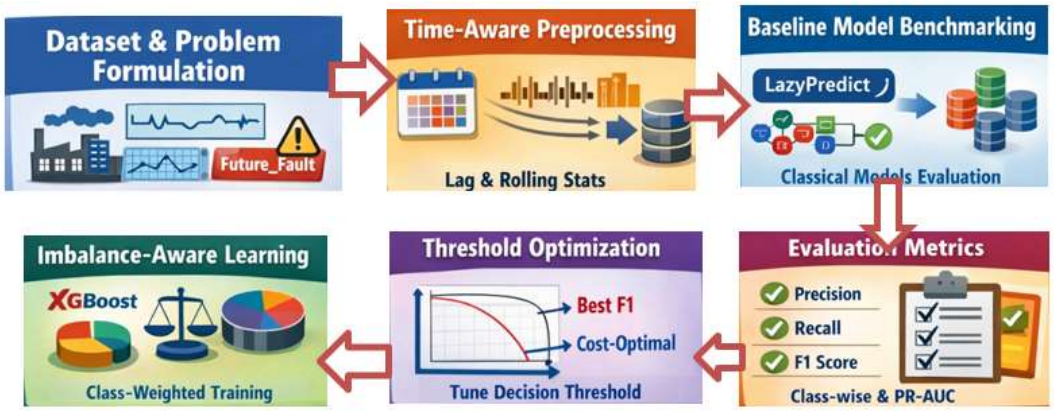

Figure 1. The proposed Methodology

The problem is redefined as an early-fault prediction task, enabling the study to be oriented towards predictive maintenance rather than fault diagnosis. Passing Future_Fault is a binary label created by predicting the allegation Future that the experiment randomly erroneously predicts in the Future, meaning that, against now, it is Future-Fault, which varies with Future. A. A fixed prediction horizon of -12 hours is used in the main experimental cases. Therefore, any fault severity that is non-zero that happens inside this horizon is considered an impending fault event. Such a change makes the model's goal consistent with the need to predict failures before they occur.

Time Aware Preprocessing and Feature Engineering

Chronology is maintained by sorting all records by the timestamp field. Instead of a standard random train-test split, a temporally stratified split is used: the first 80 per cent of the acquisitions comprise the training set, and the remaining 20 per cent comprise the test set. This will eliminate the possibility of information leaking into the Future and is more realistic for real-world deployment. The time characteristics of degradation are synthesized by using a selection of sensors and health indicators. These are lagged observations at 1, 3, 6, and 12 hours, along with rolling statistics, means, and standard deviations over a 6-hour window. These characteristics convert the real-time Digital Twin measurements into more comprehensive time-contextual representations, in which the progressive failure process is coded. The promoted rows that lack adequate historical background are removed to maintain consistency.

Baseline Model Benchmarking

A preliminary benchmark is performed using the LazyPredict framework, which tests a broad range of classical machine learning classifiers across a common set of preprocessing scenarios. This current analysis clarifies the model behaviour, and as it turns out, the majority-class domination prevails: many algorithms with high accuracy are predictors of the normal state only. The benchmark thus highlights the need for imbalance-conscious strategies and the need to redefine evaluation measures.

Imbalance-Aware Learning Strategy
Faced with a severe bias toward min-max, class imbalance is addressed by class-weighted learning rather than class-weighted duplicative oversampling, which would also disrupt the temporal organization. Tree-based ensemble models are preferred for their ability to handle nonlinear feature interactions. The models are trained using weights proportional to the inverse class frequencies, thereby encouraging sensitivity to the relatively low-frequency fault events.

Probability Calibration and Threshold Optimisation
Instead of reducing the decision threshold to 0.5, the predicted probabilities are evaluated using precision-recall curves. The operating threshold is then chosen between the highest minority-class F1-score and a cost-sensitive one, which is highly sensitive to missed faults compared to false alarms. Such calibration is the setting of model predictions to match the maintenance imperative, as it acknowledges that the cost of failed systems, as assumed over time, that are not incurable, is more than that of incurable codes of undetected failure.

Evaluation Metrics
Precision, recall, F1-score (class-wise and macro-averaged), balanced accuracy, and PR-AUC are used to evaluate model performance. Accuracy is reduced since it is prone to inaccuracy when there is class imbalance. This evaluation paradigm will ensure that reported results accurately reflect the actual early-warning capability, free of the majority-class bias.
Altogether, this study design guarantees methodological rigour by combining temporal validation, realistic labelling, imbalance-sensitive learning, and decision-focused assessment, making the proposed framework applicable for implementation in Digital Twin-based predictive maintenance systems.

RESULTS

This section outlines a systematic experimental study to explore early fault detection in an industrial context using Digital Twin technology. The experiments proceed step by step to highlight core challenges such as class imbalance, temporal dependency, and misleading accuracy and to show how each change in methodology strengthens the decision. All experiments use the same five-year hourly data set and a strict time-based train-test split, akin to the conditions in real-world deployment.

Experiment 1. Baseline Model Benchmarking (Static Classification)
What was done: The evaluation of a wide range of classical machine-learning classifiers was performed with the use of LazyPredict and instantaneous features and the original fault labels.

Observed results (Table 1):
Most of the models had a success rate of around 69%, but the confusion matrices showed that all predictions were distilled into the majority No Fault

Table 1. Baseline Model Results

Model	Accuracy	Balanced Accuracy	F1 Score
BaggingClassifier	0.68	0.2	0.57
XGBClassifier	0.69	0.2	0.57
AdaBoostClassifier	0.69	0.2	0.57
CalibratedClassifierCV	0.69	0.2	0.57
BernoulliNB	0.69	0.2	0.57
ExtraTreesClassifier	0.69	0.2	0.57
LinearSVC	0.69	0.2	0.57
DummyClassifier	0.69	0.2	0.57
SGDClassifier	0.69	0.2	0.57
RidgeClassifierCV	0.69	0.2	0.57

Key insight: High accuracy was due solely to class imbalance, indicating that fixed classification is not suitable for predictive maintenance.

Experiment 2. Reformulating the Task as Early Fault Prediction
What changed: The multiclass diagnostic problem was reposed as a binary, early-fault prediction task with a 12-hour prediction horizon, without pursuing any time-based feature engineering or balancing.

Table 2. Experiment 2 Results

Metric	Value
Accuracy	0.6946
Balanced Accuracy	0.5
Macro F1-score	0.4099
Fault Precision	0
Fault Recall	0
Fault F1-score	0
PR-AUC	0.3048

Table 2 shows that high accuracy may be attributed to the prevalence of the majority group; the fault recall measure is zero, indicating that the model does not identify any anticipated faults. This finding shows that fault prediction is significantly harder than fault diagnosis, and that reformulating the task alone is not sufficient in the presence of time and class imbalance.

Experiment 3. Time-based Feature Engineering
What changed: To identify the short-term degradation drivers, the model was further extended with lag features at 1, 3, 6, and 12 hours, and rolling statistics (mean and standard deviation), without changing the underlying classifier.

Table 3. Experiment 3 Results

Metric	Value
Accuracy	0.6945
Balanced Accuracy	0.5
Macro F1-score	0.4099
Fault Precision	0
Fault Recall	0
Fault F1-score	0
PR-AUC	0.3078

Table 3 shows that, despite a slight improvement in the precision-recall area under the curve, the classifier still only predicted the normal class. This finding supports the idea that time-related terms alone are insufficient to address the significant imbalance in class sizes when traditional learning objectives are used.

Experiment 4. Imbalance-Aware Learning
What changed: Random Forest was extended to include a class-weighted variant that, at the same time, maintains the time-based characteristics; this model imposes penalties for misclassifying faults.

Table 4. Experiment 4 Results

Metric	Value
Accuracy	0.6763
Balanced Accuracy	0.5034
Macro F1-score	0.4514
Fault Precision	0.3319
Fault Recall	0.059
Fault F1-score	0.1002
PR-AUC	0.3157

Table 4 shows that the model eventually learns with a finite number of faults, with a recall significantly below zero, thereby demonstrating the physical benefits of learning that considers imbalances. Although overall accuracy decreases slightly, macro-F1 and precision-recall AUC scores increase, confirming that the model has developed a tendency to capture early degradation rather than relying on majority-class predictions.

Experiment 5. Threshold Optimisation and Cost-Sensitive Decision Making
What changed: Instead of using the default probability cut-off of 0.5, precision-recall analysis was used to set decision thresholds, which included:
- The Best F1 threshold
- false negative cost A false negative cost is highly sensitive to cost.

Table 5. Threshold optimization and cost-sensitive decision making

Setting	Accuracy	BalancedAcc	MacroF1	Fault Precision	Fault Recall	Fault F1	PR-AUC
Default Threshold (0.50)	0.6763	0.5034	0.4514	0.3319	0.059	0.1002	0.3157
Best-F1 Threshold (0.40)	0.3063	0.5006	0.2353	0.3057	1	0.4683	0.3157
Cost-Sensitive Threshold (FN: FP = 10:1)	0.3063	0.5006	0.2353	0.3057	1	0.4683	0.3157

Table 5 shows that threshold optimization significantly affects the system's operational profile. The model can do this by reducing the decision threshold, thereby achieving flawless fault recall and preventing looming faults from being missed. An increase reduces the overall precision of the results by increasing the number of false alarms, but this is a reasonably acceptable trade-off when predictive maintenance is involved, where the false omission of a fault has significantly more serious consequences than a false alarm.

Table 6 presents a detailed overview of the gradual development of the early prediction of fault severity framework in four different experimental design layouts. The next experiment introduces a different modelling improvement, advancing the system to a robust, early-warning format that evolves into an entirely configurable, deployment-oriented design. The table lists the most noticeable performance improvements achieved at each developmental stage, along with the prevailing

constraint that remains. Such systematic comparative analysis highlights how methodological improvements gradually eliminate the pragmatic challenges facing asset management.

Table 6. Cross-Experiment Summary

Experiment	Main Gain	Key Limitation
Exp2	Early-warning formulation	No fault detection
Exp3	Temporal context added	Imbalance still dominates
Exp4	First meaningful fault detection	Low recall
Exp5	Configurable early-warning system	Higher false alarms

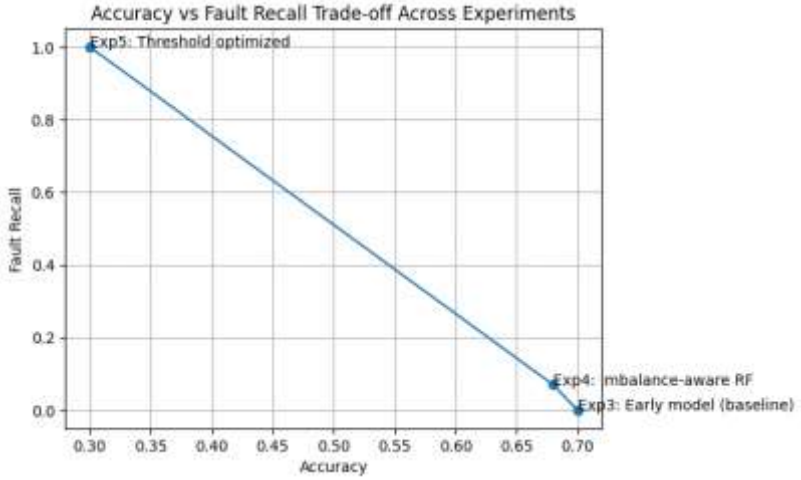

Figure 2. Accuracy vs Fault Recall Trade-off Across Experiments

Figure 2 illustrates the empirical trade-off between overall accuracy and fault recall across the series of experiments.

In Experiments 2 and 3, the models achieve remarkably high accuracy with zero fault recall; this trend indicates that the majority-class bias is dominant, suggesting that early fault-prediction ability is insufficient.

Experiment 4 shows a rather small decrease in precision and a corresponding rise in inaccuracy recall, thus reinforcing the benefits of imbalance-sensitive learning techniques.

Experiment 5 shows a deliberate change in operating point: accuracy is deliberately compromised to achieve perfect fault recall, enabling an all-alarm-sensitive predictive-maintenance regime.

All these observations show that the notion of accuracy alone, as an isolated concept, cannot reflect the performance specifications of early fault-prediction systems; recall-based decision policies are essential, particularly in the safety-critical industrial setting.

DISCUSSIONS

The outcomes of the experiments show that the early fault prediction in Digital Twin-enabled industrial systems is significantly affected by task formulation, time modelling, imbalance management, and the decision threshold. The base results of the reformulation of the initial fault-diagnosis problem into a 12-hour early fault-prediction problem are given in Table 2. Even though the overall accuracy is high (0.6946), the model does not identify any looming faults, resulting in a recall of 0 and an F1-score of 0 for the fault class. It follows that accuracy in such situations is deceptive when extreme class imbalance is present. That fault prediction in the early stages is significantly harder than fault diagnosis, with fault-related patterns being particularly weak before they occur.

The effect of temporal features engineering is reported in Table 3. When lagged variables and rolling statistics are included in the model, there is a slight improvement in PR-AUC, but fault recall remains zero. This suggests that, even though degradation dynamics require temporal context, it is inadequate when the learning objectives remain biased toward the majority class. Therefore, Table 4 compares an imbalance-sensitive Random Forest that is trained on Temporally Enriched features. This setup gives the first useful early-detection results, with a fault recall of 0.059 and a macro-F1 of 0.4514. Even though the accuracy drops to 0.6763, this trade-off favours majority-class domination over decision relevance.

Table 5 also demonstrates that the model's utility ultimately depends on the decision policy, not on the choice of classifier. Optimizing the probability threshold using precision-recall analysis yields fault recall of 1.0 and fault F1 of 0.4683, at the expense of lower accuracy. This sequence is graphically summarised in the accuracy-recall trade-off figure that indicates accuracy decreases with increasing recall in Experiments 2 and 5. In the case of predictive maintenancepredictive maintenance, such a trade-off is desirable because missing a fault is much more expensive than issuing a false alarm. On the whole, the findings suggest that reliable early fault prediction can occur only when temporal modelling, imbalance-aware learning, and threshold optimization are viewed as an interconnected set of factors that can provide a realistic and operational framework for Digital Twin-based predictive-maintenance systems.

The results in Table 6 show that a single modelling approach cannot adequately handle early fault-severity prediction. This is demonstrated by experiment 2 (Exp 2), which shows that early-warning formulations alone are invalid;

faults are completely detected when the class imbalance is extreme. In Experiment 2 (Exp 3), which adds time context, we find that representation learning is enhanced, but the influence of imbalance still prevails in predictive performance. Experiment A shows a single critical transition and achieves the first meaningful fault detection, despite limited recall due to conservative decision boundaries. Experiment 5 (Exp5): By adopting configurable decision policies, it is possible to create more sensitive early-warning systems in Experiment 5 (Exp5), which trade higher fault recall rates against higher false-alarm rates. Taken together, all these results demonstrate that efficient prediction of early severity results from a consistent combination of temporal modelling, imbalance-aware learning, and threshold calibration, rather than from individual improvements.

CONCLUSIONS

Continuous industrial telemetry in high dimensions generated by digital twin (DT) platforms offers opportunities for large-scale fault and predictive maintenance (PdM). However, the real-life applications of PdM face two lasting challenges: (i) fault events are rarely realized and tend to be imbalanced in terms of their classes, (ii) there is an urgent necessity to recognize the fault escalation at an early stage to respond to it before it goes out of proportion. To address this issue, our research proposes a deployable, exploratory study to predict faults in Digital Twin-based industrial systems at the outset using machine-learning methods. The empirical results showed that high classification accuracy alone is insufficient and not always predictive of usefulness in practice, especially in predictive maintenance environments characterized by highly skewed class distributions and strong temporal dependencies. Reconceptualizing fault diagnosis as a horizon-based early warning problem, using more advanced temporal feature engineering, and adopting imbalance-aware learning strategies enabled the proposed framework to enter a regime in which it can effectively forecast faults early in a meaningful manner. The findings highlighted that although temporal characteristics are indispensable, they alone are insufficient; nevertheless, class-weighted ensemble models were also necessary to nurture minority-class learning. The most significant was that probability threshold optimization was identified as a key mechanism for aligning the model's behaviour with operational imperatives, thereby enabling trade-offs between fault recall and false alarm rate. The last model in an alarm-sensitive setup achieved ideal fault recall, highlighting its suitability for safety-critical industrial settings where the cost of an unnoticed failure is unacceptable. Taken together, these results support the view that realistic predictive maintenance systems should have recall-oriented, decision-conscious evaluation measures rather than raw accuracy. Based on this, it is proposed that practitioners in the industry adopt a time-conscious validation strategy, imbalance-conscious training protocols, and threshold calibration when adopting algorithms for Digital Twin-based analytics. Finally, adaptive horizons, interpretable modelling, and physics-informed constraints must be incorporated into the Future, as only when these are combined can the soundness and credibility of the maintenance decision-support system in the real world be enhanced.

The contribution of this paper lies in treating early fault prediction as a deployment-oriented problem rather than a classical classification problem by integrating a time-aware component. Despite this contribution, our research has some limitations that can be considered in future research. As a sample of limitation, the experiments were conducted on a single dataset with a fixed prediction horizon.

Author Contributions: Conceptualization, O.A. ; Methodology, O.A. ; Software, O.A. ; Validation, O.A. ; Formal Analysis, O.A. ; Investigation, O.A. ; Resources, O.A. ; Data Curation, O.A. ; Writing – Original Draft Preparation, O.A. ; Writing – Review & Editing, O.A. ; Visualization, O.A. ; Supervision, O.A. ; Project Administration, O.A. ; Funding Acquisition, O.A. Authors have read and agreed to the published version of the manuscript.
Institutional Review Board Statement: Ethical review and approval were waived for this study because the research does not involve vulnerable groups or sensitive issues.
Funding: The authors received direct funding for this from Applied Science University, Manama, Kingdom of Bahrain
Acknowledgements: The author would like to thank Applied Science University, Manama, Kingdom of Bahrain, for their funding and support.
Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.
Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to restrictions.
Declaration of Generative AI and AI-Assisted Technologies in the Writing Process: During the preparation of this work, the author(s) used Grammarly for proofreading and spell checking since the Authors are not native speakers. All intellectual content, analysis, and interpretations were produced solely by the authors. After using this AI tool/service, the author(s) reviewed and edited the content as needed, taking full responsibility for the publication's content.
Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

Abd Wahab, N. H., Hasikin, K., Wee Lai, K., Xia, K., Bei, L., Huang, K., & Wu, X. (2024). Systematic review of predictive maintenance and digital twin technologies challenges, opportunities, and best practices. *PeerJ Computer Science*, 10, e1943. https://doi.org/10.7717/peerj-cs.1943

Alobaid, A., & Corcho, O. (2024). Linear approximation of the quantile–quantile plot for semantic labelling of numeric columns in tabular data. *Expert Systems with Applications*, 238, 122152. https://doi.org/10.1016/j.eswa.2023.122152

Apeiranthitis, S., Zacharia, P., Chatzopoulos, A., & Papoutsidakis, M. (2024). Predictive Maintenance of Machinery with Rotating Parts Using Convolutional Neural Networks. *Electronics*, *13*(2), 460. https://doi.org/10.3390/electronics13020460

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. D. P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & Industrial Engineering*, 137, 106024. https://doi.org/10.1016/j.cie.2019.106024

Compare, M., Baraldi, P., & Zio, E. (2020). Challenges to IoT-Enabled Predictive Maintenance for Industry 4.0. *IEEE Internet of Things Journal*, *7*(5), 4585–4597. https://doi.org/10.1109/JIOT.2019.2957029

Cook, J., & Ramadas, V. (2020). When to consult precision-recall curves. *The Stata Journal: Promoting Communications on Statistics and Stata*, *20*(1), 131–148. https://doi.org/10.1177/1536867X20909693

Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., & Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. *Computers in Industry*, 123, 103298. https://doi.org/10.1016/j.compind.2020.103298

Grieves, M., & Vickers, J. (2017). Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In F.-J. Kahlen, S. Flumerfelt, & A. Alves (Eds.), Transdisciplinary Perspectives on Complex Systems (pp. 85–113). Springer International Publishing. https://doi.org/10.1007/978-3-319-38756-7_4

Hancock, J. T., Khoshgoftaar, T. M., & Johnson, J. M. (2023). Evaluating classifier performance with highly imbalanced Big Data. *Journal of Big Data*, *10*(1), 42. https://doi.org/10.1186/s40537-023-00724-5

Hancock, J. T., Khoshgoftaar, T. M., & Johnson, J. M. (2024). Using Area Under the Precision Recall Curve to Assess the Effect of Random Undersampling in the Classification of Imbalanced Medicare Big Data. *International Journal of Reliability, Quality and Safety Engineering*, *31*(1), 2350039. https://doi.org/10.1142/S0218539323500390

IndFD-PM-DT. (2024). Retrieved January 20, 2026, from https://www.kaggle.com/datasets/datasetengineer/indfd-pm-dt

Isbilen, F., Bektas, O., & Konar, M. (2025). Deep learning and similarity-based models for predicting turbofan engine remaining useful life: Insights from the CMAPSS dataset. *The Aeronautical Journal*, *129*(1337), 2004–2035. https://doi.org/10.1017/aer.2025.25

Jaenal, A., Ruiz-Sarmiento, J.-R., & Gonzalez-Jimenez, J. (2024). MachNet, a general Deep Learning architecture for Predictive Maintenance within the Industry 4.0 paradigm. *Engineering Applications of Artificial Intelligence*, 127, 107365. https://doi.org/10.1016/j.engappai.2023.107365

Karyofyllas, G., Giagopoulos, D., Jia, X., & Papadimitriou, C. (2025). A digital twin-driven machine learning framework for structural condition monitoring using multi-datasets. *Structural Health Monitoring*, 14759217251324110. https://doi.org/10.1177/14759217251324110

Li, X., Ding, Q., & Sun, J.-Q. (2018). Remaining useful life estimation in prognostics using deep convolution neural networks. *Reliability Engineering & System Safety*, 172, 1–11. https://doi.org/10.1016/j.ress.2017.11.021

Long, T., Luo, S., Luan, X., Sun, F., & Zhou, Q. (2026). A digital twin-assisted fault diagnosis framework based on denoising diffusion probabilistic model. *Measurement*, 258, 119396. https://doi.org/10.1016/j.measurement.2025.119396

Lyu, Y., Li, H., Sayagh, M., Jiang, Z. M. (Jack), & Hassan, A. E. (2021). An Empirical Study of the Impact of Data Splitting Decisions on the Performance of AIOps Solutions. *ACM Transactions on Software Engineering and Methodology*, *30*(4), 1–38. https://doi.org/10.1145/3447876

Mateus, B. C., Mendes, M., Farinha, J. T., & Martins, A. (2025). Hybrid Deep Learning for Predictive Maintenance: LSTM, GRU, CNN, and Dense Models Applied to Transformer Failure Forecasting. *Energies*, *18*(21), 5634. https://doi.org/10.3390/en18215634

Pu, D., & Wu, Y. (2024). Error Model for the Assimilation of All-Sky FY-4A/AGRI Infrared Radiance Observations. *Sensors*, *24*(8), 2572. https://doi.org/10.3390/s24082572

Richardson, E., Trevizani, R., Greenbaum, J. A., Carter, H., Nielsen, M., & Peters, B. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns*, *5*(6), 100994. https://doi.org/10.1016/j.patter.2024.100994

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Semeraro, C., Lezoche, M., Panetto, H., & Dassisti, M. (2021). Digital twin paradigm: A systematic literature review. Computers in Industry, 130, 103469. https://doi.org/10.1016/j.compind.2021.103469

Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022). Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects. *Applied Intelligence*, *52*(10), 10934–10964. https://doi.org/10.1007/s10489-021-03004-y

Sofaer, H. R., Hoeting, J. A., & Jarnevich, C. S. (2019). The area under the precision-recall curve as a performance metric for rare binary events. *Methods in Ecology and Evolution*, *10*(4), 565–577. https://doi.org/10.1111/2041-210X.13140

Tan, J., Radhi, R. M., Shirini, K., Gharehveran, S. S., Parisooz, Z., Khosravi, M., & Azarinfar, H. (2025). Innovative framework for fault detection and system resilience in hydropower operations using digital twins and deep learning. *Scientific Reports*, *15*(1), 15669. https://doi.org/10.1038/s41598-025-98235-1

Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital Twin in Industry: State-of-the-Art. *IEEE Transactions on Industrial Informatics*, *15*(4), 2405–2415. https://doi.org/10.1109/TII.2018.2873186

Yang, C., Cai, B., Zhang, R., Zou, Z., Kong, X., Shao, X., Liu, Y., Shao, H., & Akbar Khan, J. (2023). Cross-validation enhanced digital twin driven fault diagnosis methodology for minor faults of subsea production control system. *Mechanical Systems and Signal Processing*, 204, 110813. https://doi.org/10.1016/j.ymssp.2023.110813

Yoo, J. (2025). Enhancing Nickel Matte Grade Prediction Using SMOTE-Based Data Augmentation and Stacking Ensemble Learning for Limited Dataset. *Processes*, *13*(3), 754. https://doi.org/10.3390/pr13030754

You, Y., Chen, C., Hu, F., Liu, Y., & Ji, Z. (2022). Advances of Digital Twins for Predictive Maintenance. *Procedia Computer Science*, 200, 1471–1480. https://doi.org/10.1016/j.procs.2022.01.348

Zayed, S. M., Attiya, G., El-Sayed, A., Sayed, A., & Hemdan, E. E.-D. (2023). An Efficient Fault Diagnosis Framework for Digital Twins Using Optimized Machine Learning Models in Smart Industrial Control Systems. *International Journal of Computational Intelligence Systems*, *16*(1), 69. https://doi.org/10.1007/s44196-023-00241-6

Zhao, P., Luo, C., Qiao, B., Wang, L., Rajmohan, S., Lin, Q., & Zhang, D. (2022). T-SMOTE: Temporal-oriented Synthetic Minority Oversampling Technique for Imbalanced Time Series Classification. Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, 2406–2412. https://doi.org/10.24963/ijcai.2022/334

Zhong, D., Xia, Z., Zhu, Y., & Duan, J. (2023). Overview of predictive maintenance based on digital twin technology. *Heliyon*, *9*(4), e14534. https://doi.org/10.1016/j.heliyon.2023.e14534

Zonta, T., Da Costa, C. A., Da Rosa Righi, R., De Lima, M. J., Da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. *Computers & Industrial Engineering*, 150, 106889. https://doi.org/10.1016/j.cie.2020.106889