International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

HYBRID DEEP LEARNING MODELS FOR EQUIPMENT FAILURE PREDICTION IN U.S. INDUSTRIAL SYSTEMS

Mahfuz Alam[1], Sanjib Kumar Shil[2], Farmina Sharmin[3], Aashish K C[4], Abu Hena Md Martuza Ali[5], Kazi Md Shahadat Hossain[6], Abdur Rahim[7] and Sirapa Malla[8]

[1]MBA in Business Analytics, International American University

[2]MBA in Management Information System, International American University

[3]MBA in Business Analytics, International American University, Los Angeles Main Campus

[4]Master of Science in Computer and Information Science(Software Engineering), Gannon University, Erie, Pa

[5]Master's in Strategic Communication, Gannon University, Erie, Pa

[6]Master of Business Administration in Logistics Management, Central Michigan University, Mount Pleasant, Michigan, USA

[7]Master of Science in Computer Science, University of New Haven.

[8]Master's in Data Science, Gannon University, Erie, Pa

Corresponding Author: Mahfuz Alam, **Email:** jnmahfuz@gmail.com

Abstract

Equipment failure prediction plays an important role in U.S. industrial systems, where unexpected downtime carries heavy economic and safety consequences. Deep learning methods are now widely promoted for predictive maintenance, yet empirical evidence showing that hybrid architectures consistently outperform strong classical models under realistic, leakage-safe evaluation remains limited. This study examines early failure prediction using the NASA C-MAPSS turbofan engine dataset and frames the task as a binary classification problem that flags failure-imminent conditions within a fixed prediction horizon. The analysis relies on a rigorous experimental pipeline that includes engine-level data partitioning, sliding-window temporal representations, and carefully defined failure labels. Classical baseline models built on engineered statistical features are evaluated alongside several deep learning architectures, including LSTM, CNN, CNN–LSTM hybrids, and LSTM models enhanced with attention mechanisms. A series of ablation studies explores the practical value of architectural hybridity, the influence of temporal window length and prediction horizon, and sensitivity to sensor removal. The results show that classical models, with gradient boosting as a notable example, deliver very strong performance with a healthy balance between precision and recall. Deep learning models reach comparably high ROC-AUC values, yet their recall for imminent failures drops sharply under standard decision thresholds. The ablation findings further reveal that hybrid architectures do not consistently outperform simpler designs and that performance depends strongly on temporal configuration and sensor choice. Taken together, these results indicate that hybrid deep learning models do not automatically earn their added complexity for

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
equipment failure prediction. The study reinforces the value of strong baselines, transparent evaluation practices, and decision-oriented metrics in predictive maintenance research.

Keywords: Predictive Maintenance, Equipment Failure Prediction, Hybrid Deep Learning, LSTM, CNN, C-MAPSS, Ablation Study

1. Introduction

1.1 Background and Motivation

Predictive maintenance now sits at the heart of many modern industrial operations, particularly in safety-critical fields such as aviation, manufacturing, and energy. In these environments, an unexpected equipment failure is not a minor inconvenience. It can escalate into serious safety concerns and significant financial losses. Maintenance strategies based on fixed schedules or reactive repairs are increasingly hard to justify in systems where sensors continuously record detailed operational data. As a result, data-driven predictive maintenance has gained momentum as a way to anticipate failures earlier and support maintenance decisions based on actual equipment condition. Surveys by Carvalho et al. (2019) and Lei et al. (2018) make a consistent point: effective predictive maintenance depends on extracting meaningful degradation patterns from high-dimensional, multivariate sensor data and translating those patterns into failure predictions that engineers can act on [3][12].

Against this backdrop, machine learning has become a natural fit for predictive maintenance problems. These methods are well-suited to capturing nonlinear relationships between sensor readings and underlying system health. Much of the early work in the field relied on careful feature engineering combined with classical models, and these approaches often delivered strong performance while remaining interpretable and relatively efficient to deploy. Guo et al. (2017) show that even relatively simple recurrent architectures can produce effective health indicators for remaining useful life prediction when the underlying signals are of high quality [8]. Zhang et al. (2019) further note that many industrial predictive maintenance problems are already well structured, meaning that increased model complexity does not automatically translate into meaningful performance gains [22]. These findings point toward the need to question the assumption that more complex models are always the right choice in this domain.

At the same time, recent years have brought growing interest in deep and hybrid learning architectures, often motivated by their capacity to model long-term temporal dependencies and subtle degradation dynamics. Similar patterns appear in other safety-critical predictive domains. Das et al. (2025), working on AI-driven cybersecurity threat detection, show that both simple and complex models can be effective depending on data characteristics and operational constraints, reinforcing the view that model selection should be guided by empirical benefit rather than architectural novelty [5]. Debnath et al. (2025), in their study of renewable energy cybersecurity, emphasize that robustness and reliability often matter more than marginal gains in accuracy in critical infrastructure settings [7]. These insights align closely with predictive maintenance, where missed failures tend to carry outsized consequences.

The broader industrial ecosystem reinforces the importance of dependable and well-calibrated predictive models. Hasan et al. (2025) show that supply chain risk management systems in the United States increasingly rely on predictive analytics, with effectiveness shaped not only by

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
accuracy but also by stability under changing conditions and interpretability for decision-makers [9]. Viewed together, this body of work suggests that the main challenge in predictive maintenance is not a lack of powerful models. The challenge lies in the absence of rigorous evaluation frameworks that can determine when added complexity delivers real value. This study is motivated by the need to critically examine whether hybrid deep learning architectures provide functional advantages for early failure prediction, or whether carefully designed baselines already capture most of the available signal.

1.2 Problem Statement

Although predictive maintenance has been studied extensively, the literature shows recurring methodological weaknesses that make meaningful comparisons between modeling approaches difficult. One common issue is the limited use of strong classical baselines. Many studies introduce deep or hybrid models without showing that these approaches outperform simpler alternatives under identical experimental conditions, even though prior work demonstrates that feature-based models can perform very well [8][22]. This practice makes it unclear whether reported improvements arise from genuine modeling advances or from choices in experimental design. Another major concern involves data leakage, particularly in time-series datasets that span multiple units or machines. Inappropriate splitting strategies, such as random sampling across timesteps, can inflate performance estimates by allowing information from the same physical system to appear in both training and test sets. Lei et al. (2018) warn that this kind of leakage undermines the validity of remaining useful life and failure prediction studies because it violates the assumption that training and evaluation data are independent [12]. Even so, this issue continues to appear across applied predictive maintenance research [3][22].

A further limitation is the lack of systematic ablation studies to justify architectural complexity. Hybrid models are often proposed without isolating the contribution of individual components, leaving unanswered questions about whether additional layers or mechanisms truly improve performance. Shivogo (2025), writing on credit scoring under concept drift, argues that complex and opaque models can become unstable when data distributions change, making strong baselines and stress testing essential in high-stakes decision systems [18]. The same reasoning applies to equipment failure prediction, where operating conditions and degradation pathways can vary across engines and over time. At the same time, growing use of predictive analytics for early warning systems in other fields highlights the importance of decision-oriented evaluation. Chouksey et al. (2025) demonstrate that early warning models for financial risk in the U.S. digital economy should be evaluated not only on discrimination metrics but also on their ability to support timely and reliable interventions [4]. By extension, early failure prediction models should be evaluated based on how well they identify impending failures with enough lead time and acceptable false alarm rates. The problem addressed in this work is whether hybrid deep learning architectures meaningfully improve early failure prediction performance when measured against strong baselines, evaluated under leakage-safe protocols, and examined with explicit attention to their structural assumptions.

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

1.3 Research Objectives and Contributions

This study aims to go past routine model comparisons and offer a careful, evidence-based look at early failure prediction methods. The central goal is to frame early failure prediction as a decision-focused task. The emphasis is on reliably flagging conditions where failure is close within a usable prediction window, not on producing abstract rankings of engine health. This perspective mirrors how such systems are used in practice, where maintenance planning depends on timely signals that support proactive action. A second objective focuses on building and evaluating strong classical baselines grounded in interpretable, domain-informed features. By showing what is possible with simpler models and thoughtfully designed inputs, the study establishes a demanding reference point. Any added complexity must demonstrate clear value beyond this foundation. This focus on baselines also helps separate gains driven by modeling choices from those driven by data handling or experimental setup. The third objective is a systematic evaluation of deep learning architectures, including recurrent, convolutional, hybrid, and attention-based models, under tightly controlled conditions. All models follow the same engine-level data splits to avoid leakage, the same preprocessing steps, and the same optimization procedures. This design ensures that performance differences can be traced to modeling capacity and inductive bias, not implementation quirks.

A fourth objective involves extensive ablation experiments that probe architectural decisions, temporal context length, and sensor selection. By selectively removing or modifying components, the study examines whether hybrid structures and deeper temporal modeling play a meaningful functional role. This analysis is tied to practical concerns around computational cost and resource usage, acknowledging that more complex models increase training and deployment demands and therefore need clear empirical justification. The contributions of this work include a rigorous experimental pipeline for early failure prediction, a clear and transparent comparison between classical and deep learning approaches, and evidence-based insights into how far architectural complexity actually helps in this setting. The study does not presume that hybrid deep models are superior. It asks whether their added complexity is supported by measurable improvements in performance that matter for real decisions.

2. Literature Review

2.1 Predictive Maintenance and Equipment Failure Modeling

Predictive maintenance has gradually moved away from traditional reliability engineering and toward data-driven methods as industrial systems have become saturated with sensors. Early work on equipment failure modeling leaned on physics-based degradation models, survival analysis, and rule-based condition monitoring. In these settings, alarms fired when signals crossed predefined thresholds. The reasoning behind these approaches was straightforward and closely tied to physical intuition, which made them appealing and easy to trust. Over time, their practical value weakened as systems grew more complex, components began interacting in subtle ways, and sensor readings became noisier and higher in dimension. As industrial datasets expanded in scale and variety, supervised learning offered a workable way to deal with this rising complexity. These models learned degradation patterns directly from historical operating data, without depending on manually defined rules. Zheng et al. (2017) showed that recurrent neural networks, particularly LSTM models, could estimate remaining useful life by capturing

Received: August 08, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
long-range temporal structure in sensor streams. This work helped move the field away from static health indicators and toward sequence-based modeling [23]. Along similar lines, Yuan et al. (2016) demonstrated that LSTM-based approaches could outperform traditional techniques for aero-engine fault diagnosis and life estimation by learning nonlinear degradation paths from time-series data [21].

Even with these developments, the predictive maintenance literature repeatedly emphasizes that performance must be assessed under conditions that reflect real industrial use. Shawon et al. (2025), in their study on improving supply chain resilience across U.S. regions, note that predictive systems in industrial environments often operate with limited data, shifting operating regimes, and high costs when predictions fail. They argue that evaluation should go beyond headline accuracy metrics [17]. The same logic applies to equipment failure modeling. Missed failures can result in serious damage or safety incidents, and frequent false alarms lead to unnecessary maintenance and reduced availability. Effective models need to flag early signs of degradation while remaining stable in the presence of noise and routine variation. The move toward data-driven predictive maintenance has brought an ongoing balance challenge between expressive modeling capacity and dependable behavior in practice. Deep learning methods offer powerful ways to represent complex patterns, yet their use in safety-critical settings remains shaped by questions of interpretability, calibration, and generalization beyond training conditions. Reviews of the predictive maintenance literature often point out that many studies focus on improving predictive scores without closely examining how models behave across different stages of an asset's life or under changing degradation patterns. This open issue motivates further work on evaluation strategies, especially when the primary goal is early failure detection instead of retrospective diagnosis.

2.2 Classical Machine Learning for Failure Prediction

Traditional machine learning paradigms remain the foundational pillars of prognostic maintenance, especially within frameworks where feature engineering can effectively encapsulate the nuances of systemic deterioration. Paredes et al. (2025) describe a hybrid framework for industrial predictive maintenance that pairs thoughtful feature engineering with ensemble learning, showing that tree-based models can deliver strong and stable performance across a range of operating conditions [13]. Their work reinforces the idea that classical approaches deserve to be treated as serious contenders rather than placeholders added for completeness. Evidence from related domains points in the same direction. Reza et al. (2025), studying socioeconomic data for U.S. citizens, find that tree-based ensembles and gradient boosting models often perform on par with, or better than, more elaborate methods on structured tabular data, especially when stability and interpretability are priorities [16]. This observation transfers naturally to predictive maintenance datasets such as C-MAPSS, where features derived from sliding windows of sensor readings often contain a clear and well-organized predictive signal. Even so, many studies in this area omit strong classical baselines or discuss them only briefly, which can inflate the perceived advantage of deep learning models.

Classical approaches also bring practical benefits tied to transparency and calibration. Logistic regression yields probabilistic outputs that are straightforward to interpret, and ensemble tree

Received: August 08, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
models provide feature importance measures that support sensor selection and robustness analysis. In safety-critical settings, these characteristics function as essential requirements rather than optional conveniences. Taken together, the literature suggests that any assessment of advanced architectures should begin by establishing whether carefully designed classical models already satisfy operational needs. Skipping this step risks confusing novelty in model design with genuine gains in predictive performance.

2.3 Deep Learning for Time-Series Degradation Modeling

Deep learning has taken center stage in time-series degradation modeling for predictive maintenance, largely because it can learn useful representations directly from raw sensor data. Early work by Babu et al. (2016) introduced deep convolutional neural networks for remaining useful life estimation and showed that convolutional filters can pick up degradation patterns from multivariate sensor streams without hand-crafted features [2]. That work made a simple point clear. Local temporal patterns carry meaningful information about system health, especially when degradation unfolds gradually and does not show up as a sharp fault signal. Recurrent architectures built on this idea by modeling temporal dependence explicitly. Studies by Zheng et al. (2017) and Yuan et al. (2016) report that LSTM-based models outperform traditional approaches by capturing long-range structure in engine degradation trajectories [23][21]. These findings helped shape the view that sequence modeling matters for high-quality prognostics. At the same time, later work has drawn attention to how deep learning results are often presented. Many papers focus on ROC-AUC or regression error metrics without examining recall at thresholds that matter for decisions. That practice can hide weak early failure detection behavior even when ranking performance looks strong.

Deep learning models also react strongly to choices around training setup, data splitting, and class balance. If engine-level separation is not enforced or normalization is handled carelessly, models may latch onto engine-specific signatures instead of learning degradation patterns that generalize. Reported performance can then look better than it should. Taken together, the literature suggests that deep learning offers powerful tools for modeling time-series degradation, though the gains depend heavily on careful experimental design and evaluation that reflect operational goals.

2.4 Hybrid Architectures and Attention Mechanisms

Hybrid architectures that combine convolutional and recurrent components have been proposed to capture both short-term and long-term aspects of degradation. Peng et al. (2022) present a spatio-temporal attention-based framework for turbofan engine remaining useful life prediction, arguing that attention mechanisms allow models to focus on informative time steps and sensor interactions [14]. De Luca et al. (2023) apply attention-based deep models to the NASA turbofan dataset and report improvements over baseline recurrent architectures [6]. These studies suggest that hybrid designs and attention may help models emphasize critical phases of degradation.

The empirical case for this added complexity often remains thin. Many papers introduce hybrid structures or attention layers without running ablation studies that isolate what each component

Received: August 08, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
contributes. When that analysis is missing, it becomes difficult to tell whether performance gains come from architectural ideas or from increased parameter counts and training flexibility. Similar concerns appear in other high-stakes areas. Shivogo (2025), studying credit scoring under concept drift, shows that complex hybrid models can behave unpredictably when data distributions shift, reinforcing the need for stress testing and component-level scrutiny [18]. In predictive maintenance, operating conditions and degradation paths can differ substantially across engines. Without systematic ablation, confidence in hybrid architectures stays limited. Attention mechanisms are often described as tools for interpretability, yet their correspondence to physical degradation processes is rarely examined in depth. The literature points toward a need for more disciplined evaluation of hybrid models, with a focus on whether extra components deliver practical benefits under realistic conditions.

2.5 Research Gaps

Across the predictive maintenance literature, several gaps continue to show up. First, there is heavy reliance on deep learning models without careful benchmarking against strong classical baselines. Work by Paredes et al. (2025), supported by results from other structured prediction settings, shows that classical models remain highly competitive when they are thoughtfully designed [13][16]. Second, sensitivity to prediction horizon plus temporal context receives limited attention, despite its importance for early failure detection. Many models look impressive close to end-of-life yet fail to provide usable lead time, a shortcoming that is rarely examined directly.

Third, sensor importance plus robustness analysis receives limited focus. Many studies report aggregate performance metrics while leaving unanswered questions about model behavior when sensors are removed, corrupted, or reordered. Hasan et al. (2025) identify related weaknesses in explainable AI systems for supplier credit approval, pointing out that high-stakes decision systems often lack robustness plus transparency analysis even when real-world impact is significant [10]. Finally, calibration plus threshold selection receives limited emphasis in predictive maintenance research. Shivogo (2025) argues that attention limited to ranking metrics hides decision risk plus instability, a concern that maps directly to early failure prediction, where poor calibration can carry serious consequences [18]. These gaps motivate the study's focus on strong baselines, leakage-safe experimental design, ablation studies, plus decision-oriented evaluation. By addressing these issues directly, this work aims to offer a more grounded assessment of hybrid deep learning models for equipment failure prediction that reflects real operational demands.

3. Methodology

3.1 Dataset Description

This study uses the NASA Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) FD001 dataset, a well-known reference point in predictive maintenance research. The dataset contains multivariate time-series data generated from a detailed simulation of turbofan engine degradation under controlled operating conditions. Each engine trajectory covers a complete operational lifecycle that ends in a failure event, which supports supervised learning for failure prediction. The FD001 subset includes data from 63 distinct engines, with lifecycle

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
lengths ranging from 8 to 287 operational cycles, plus an average length of roughly 195 cycles. For every engine cycle, the dataset provides three operational settings along with 21 sensor measurements that describe different aspects of engine behavior, such as temperature, pressure, plus rotational speed. An initial statistical review was carried out to evaluate data quality plus sensor usefulness. Ten sensor channels contained only missing values, so they were removed. Several other sensors showed near-zero variance across all engine lifecycles, which suggested minimal value for modeling degradation patterns. These low-variance sensors were identified using a standard deviation threshold, then excluded from further analysis. After this filtering step, eight sensor channels showing meaningful variability plus degradation behavior were kept for modeling. This choice helped keep attention on informative signals without introducing noise from features that contribute little.

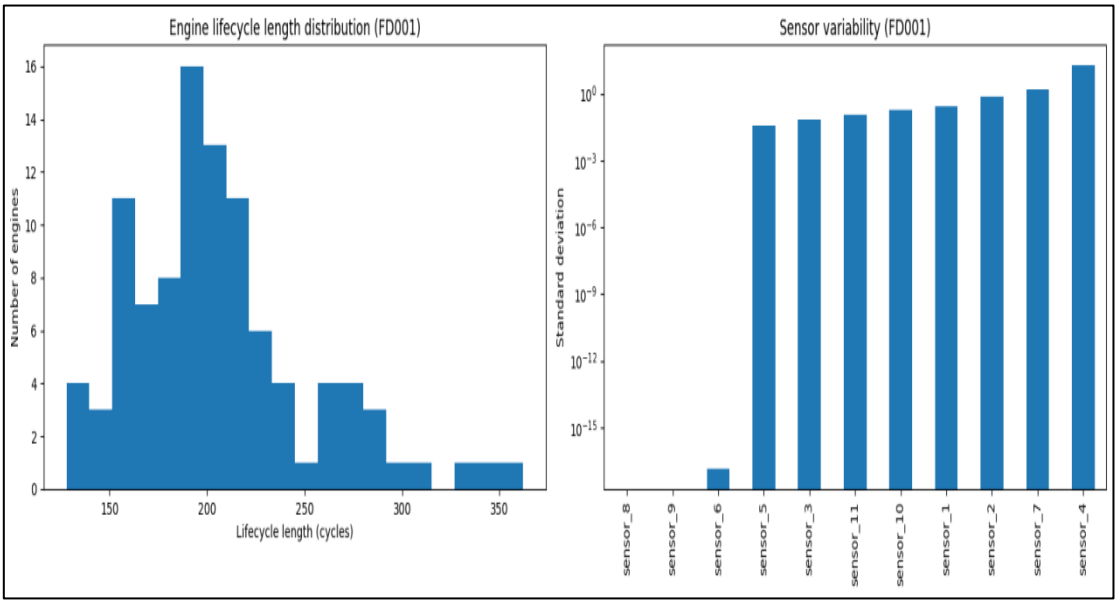

Fig.1: Engine lifecycle and sensor variability for the FD001 subset

3.2 Problem Formulation

The predictive maintenance task is set up as a binary early failure prediction problem. The goal is not to estimate the exact remaining useful life of an engine. The objective is to decide whether an engine is likely to fail within a defined future window. This framing matches real maintenance settings, where the key question is whether action is needed soon. Remaining Useful Life (RUL) is calculated for each engine cycle as the difference between the final failure cycle plus the current cycle index. Each cycle then receives a binary label based on a chosen prediction horizon. A cycle is labeled as failure imminent if its RUL is less than or equal to 20 cycles; otherwise, it is labeled as normal operation. This horizon reflects a practical planning window for maintenance, offering enough lead time for inspection or repair actions without making the prediction task trivial. Model inputs are built using sliding windows of consecutive sensor readings. Each sample includes sensor measurements from the previous 30 cycles, which are used to predict whether failure will occur within the next 20 cycles. This window-based setup allows models to learn short-term fluctuations together with longer-term degradation patterns, closely matching how condition monitoring systems are used in industrial environments.

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

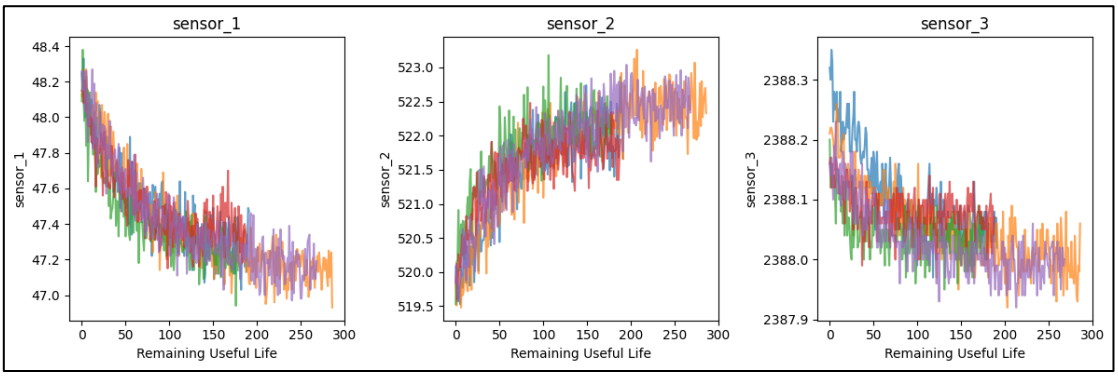

Fig.2: Remaining useful life across sample sensors

3.3 Data Preprocessing

The preprocessing pipeline is designed to support fair evaluation while removing sources of information leakage that could inflate performance estimates. It begins with the removal of invalid sensor channels plus near-constant features, as outlined in Section 3.1. The remaining sensor readings keep their original temporal order within each engine trajectory. Remaining Useful Life values are computed separately for each engine, which prevents RUL calculations from using information from other engines. A sliding window method is then applied to the sensor data. For each engine, overlapping windows of fixed length are extracted, with each window treated as an independent sample. The associated binary failure label is assigned based on the RUL at the end of the prediction horizon relative to the window endpoint.

To avoid data leakage, the dataset is split at the engine level instead of the window level. Engines are randomly assigned to training or test sets using an 80/20 split, with all windows from a single engine placed entirely within one split. This ensures that evaluation uses completely unseen engine units, preventing models from exploiting similarities across windows from the same engine. Feature normalization relies only on statistics from the training set. For each sensor channel, the mean plus standard deviation is calculated across all training windows, then used to standardize both training plus test data. This prevents test-set information from influencing normalization. No resampling or class balancing methods are used, leaving the natural class imbalance intact to reflect real operational settings where failures occur infrequently.

3.4 Exploratory Analysis Focused on Failure Signals

The exploratory analysis was shaped around one goal: identifying degradation patterns that carry physical meaning plus real predictive value, rather than producing surface-level visual summaries. Since the aim is early failure prediction, the focus stayed on how sensor behavior changes as engines move toward end-of-life, with particular attention to signals that show consistent trends, steady progression, or clear structural shifts under sustained degradation. The first step looked at individual sensor trajectories aligned by Remaining Useful Life rather than absolute cycle count. This simple change made a noticeable difference. Patterns that appeared noisy or ambiguous in raw cycle-indexed plots became much clearer once aligned by RUL. Several sensors, including sensor_1, sensor_2, sensor_3, sensor_4, sensor_5, sensor_7,

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
sensor_10, plus sensor_11, showed systematic behavior as RUL approached zero. Some exhibited slow drifts over time, others showed nonlinear acceleration or growing volatility close to failure. What stood out was the consistency of these patterns across engines with very different lifespans. This consistency suggests these sensors are responding to shared physical wear processes, such as heat stress buildup, efficiency degradation, or mechanical imbalance, rather than short-term operational fluctuations.

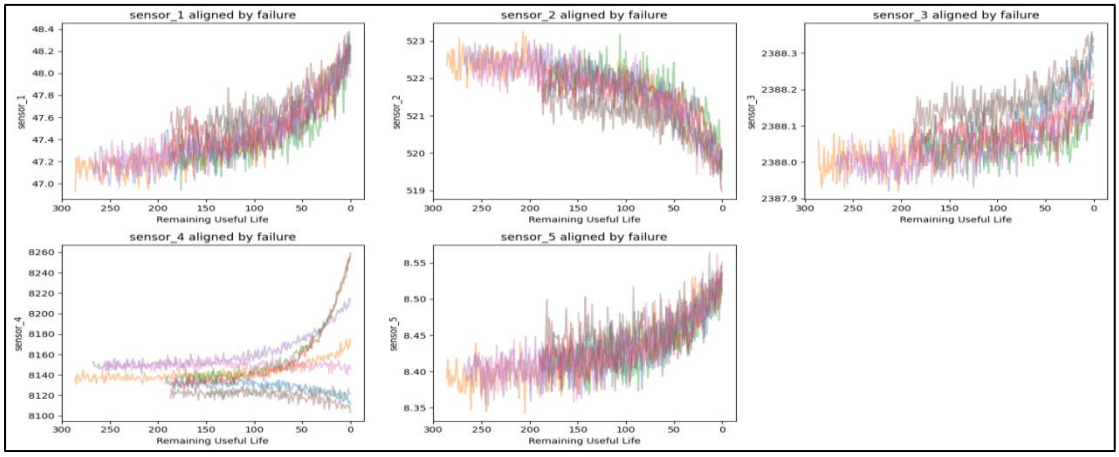

Fig.3: Remaining useful life across select sensors aligned by failure

The analysis then shifted from individual trajectories to how sensor distributions change across broad stages of the engine lifecycle. Engine cycles were grouped into coarse RUL bins corresponding to early life, mid-life, plus late life operation. Across these stages, several sensors showed clear shifts in mean values, reflecting gradual but persistent changes in operating conditions as degradation progresses. At the same time, many sensors displayed growing variance as failure drew closer, pointing to increasing instability plus shrinking control margins. This pattern fits well with how physical systems behave near their limits, where small disturbances trigger larger responses. Not every sensor followed this pattern, which reinforced the idea that only a subset truly tracks degradation. These distributional changes supported the use of window-level statistical features for baseline models, while also setting reasonable expectations that temporal models could learn from evolving dynamics inside each window.

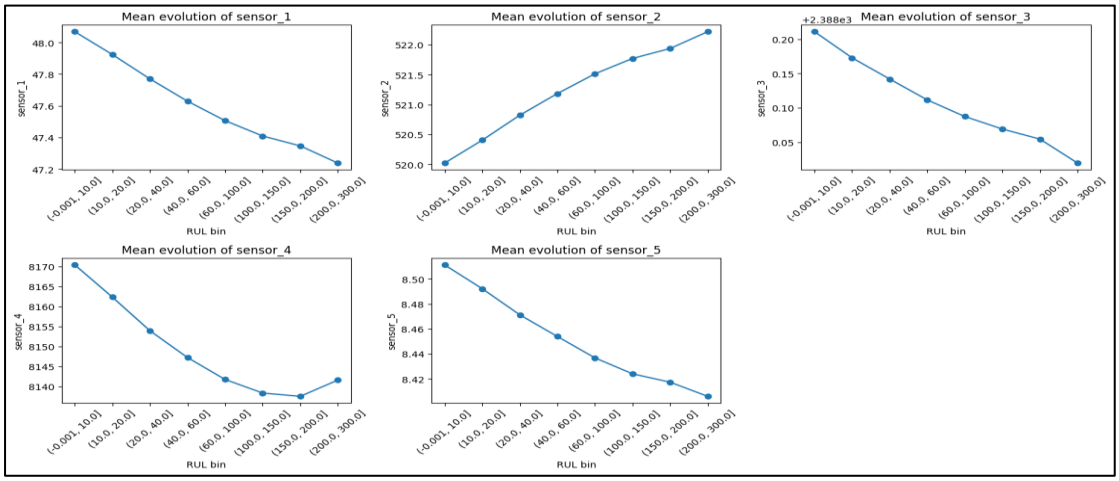

Fig.4: Mean evolution across sensors

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A multivariate view added another layer to the analysis. Inter-sensor relationships were examined across the engine lifecycle by computing correlation matrices separately for early-life plus late-life phases. The results showed clear restructuring of sensor dependencies as engines approached failure. Sensors such as sensor_4, sensor_7, plus sensor_1 exhibited the largest changes in correlation strength with other sensors. This shift suggests that interactions between components evolve during degradation, likely through cascading effects where stress or failure in one subsystem alters how others operate. These changing relationships underline why multivariate modeling matters, since single-sensor trends alone miss part of the system behavior leading up to failure.

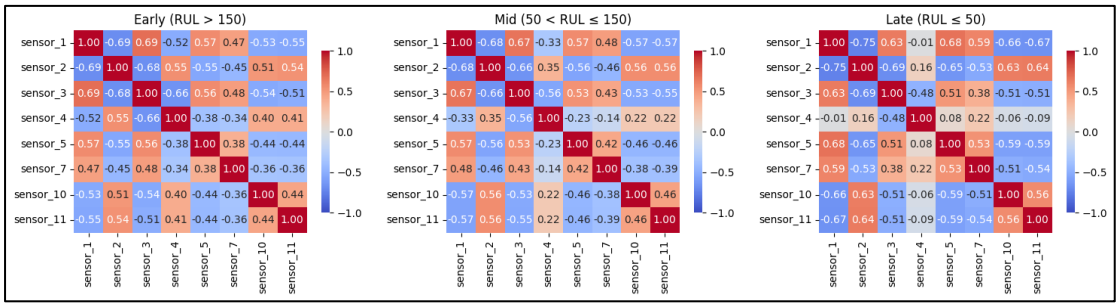

Fig.5: Inter-sensor correlations across RUL phases

Pulling these observations together, a heuristic sensor selection strategy was used to identify the most informative subset of sensors for later modeling. Sensors were ranked based on how much their mean values shifted across RUL bins, how strongly their variance increased near failure, plus how much their correlation structure changed between early plus late life. Looking across these criteria helped ensure that selected sensors carried consistent degradation signals from multiple analytical angles. The final set of eight sensors showed strong, interpretable failure-related behavior, offering a solid basis for reducing dimensionality. This step trimmed noise plus redundancy while preserving the core degradation information, which in turn supported more stable plus interpretable model training.

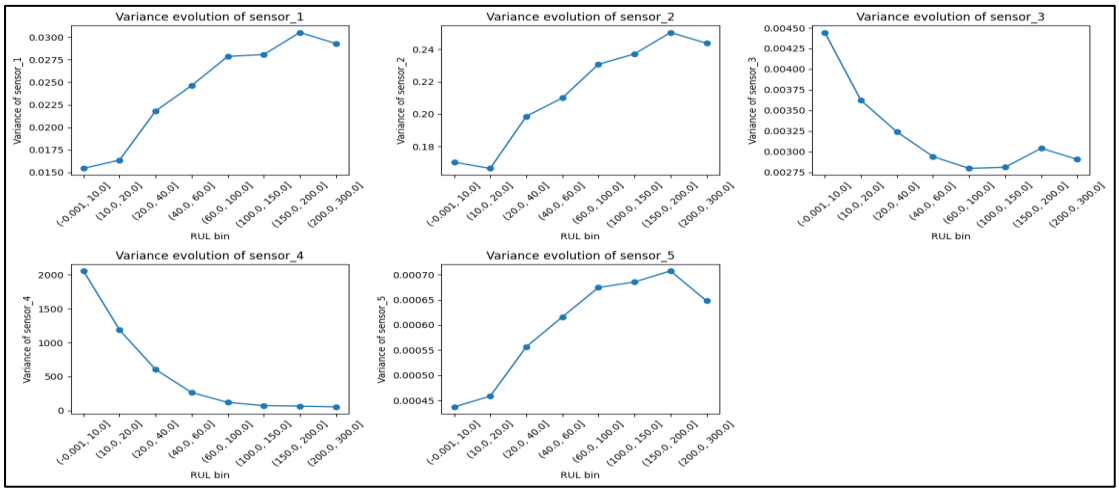

Fig.6: Variance evolution across sensors

The exploratory analysis shows that the C-MAPSS dataset contains rich, structured degradation signals that intensify as failure approaches. These signals are not evenly distributed across

Received: August 08, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
sensors, nor do they remain static over time. They evolve both temporally plus across sensor relationships. These insights directly shaped the modeling choices in this study, including the use of sliding temporal windows, the comparison between feature-based plus raw-sequence models, plus the design of sensor ablation experiments used to probe model robustness.

3.5 Feature Engineering for Baseline Models

For classical machine learning models, explicit feature engineering was used to turn raw sensor windows into compact, interpretable representations. Each sliding window spanning 30 cycles was handled on its own. For every retained sensor within a window, three summary statistics were extracted: the mean, the standard deviation, plus a temporal trend slope. The slope was calculated using simple linear regression over the sensor values inside the window, giving a rough but useful sense of how the signal was moving over time. These features were chosen to strike a practical balance between richness plus interpretability. The mean reflects the typical operating level of a sensor within the window. The standard deviation captures short-term variability plus signs of instability. The slope represents gradual degradation or recovery trends. Taken together, these statistics describe both overall behavior plus the dynamics of the sensor signals without forcing the models to infer temporal structure on their own. This mirrors common industrial practice, where engineers often prefer aggregated indicators over raw, high-frequency measurements. Feature extraction was applied in the same way to training plus test windows after splitting the data at the engine level, which ensured that no information from test engines influenced how features were built. The resulting feature vectors were standardized using statistics computed from the training set before model fitting. This feature engineering setup provides a solid, interpretable baseline against which more complex deep learning models can be evaluated in a meaningful way.

3.6 Baseline Models

To create a clear performance reference, several classical supervised learning models were implemented using the engineered statistical features. These baselines were chosen because of their broad use in industrial predictive maintenance, plus because they rely on different modeling assumptions. Logistic Regression was used as a linear probabilistic classifier, offering a transparent baseline with feature weights that are easy to inspect. Even with its simplicity, logistic regression can perform surprisingly well when the input features are informative, which makes it an important point of reference when judging whether added model complexity is justified. Random Forest classifiers were included to capture nonlinear relationships plus interactions between features without heavy parameter tuning. By combining decisions from many trees, random forests tend to be robust to noise, while also producing variable importance measures that support interpretation.

Gradient Boosting classifiers were added as a strong ensemble baseline, well known for modeling complex nonlinear patterns through sequential tree construction. In many tabular-data settings, gradient boosting approaches represent a practical upper bound on performance, which makes them a demanding benchmark for deep learning models. Alongside these supervised methods, a One-Class Support Vector Machine was implemented as an unsupervised anomaly detection approach. The OCSVM was trained using data from normal operation only, then evaluated on its ability to flag abnormal behavior close to failure. This

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
model was not meant to compete directly with supervised classifiers. Its role was to act as a sanity check on whether failure-related behavior separates cleanly from nominal operation within the feature space. All baseline models were trained plus evaluated using the same train-test splits plus the same evaluation metrics, ensuring a fair comparison across methods.

3.7 Deep Learning Architectures

The deep learning models were trained directly on raw sensor windows, keeping the original temporal structure intact. Each input sample was represented as a two-dimensional tensor, with one dimension for the window length and the other for the number of selected sensors. Before training, sensor values were normalized using statistics calculated from the training set alone. Several architectures were explored to capture temporal dependencies in different ways. A Long Short-Term Memory network was used as a temporal baseline. Its gated structure is well-suited for learning long-range dependencies and gradual degradation patterns that unfold across a window. LSTM models handle sequential data effectively, though they can be less sensitive to short, localized patterns without added structure. A Convolutional Neural Network was then applied to focus on local temporal behavior. One-dimensional convolutional filters were used along the time axis to detect short-term motifs such as abrupt changes or brief oscillations in sensor readings. These localized patterns may serve as early signals of emerging failure conditions.

To bring these perspectives together, a hybrid CNN plus LSTM architecture was designed. In this setup, convolutional layers first extract short-term temporal features, which are then passed to an LSTM layer that models longer-term dependencies and overall degradation trajectories. This design follows a common view in predictive maintenance research that failure behavior reflects an interplay between local anomalies and cumulative wear. An additional variant extended the LSTM architecture with an attention mechanism. This model was used to examine whether selectively weighting different time steps within a window improves prediction performance. Attention enables the model to focus more strongly on segments of the sequence that carry the most information for failure prediction, with potential gains in robustness and interpretability. All deep learning models were trained using early stopping based on validation recall, reflecting the higher operational cost associated with missed failures compared to false alarms.

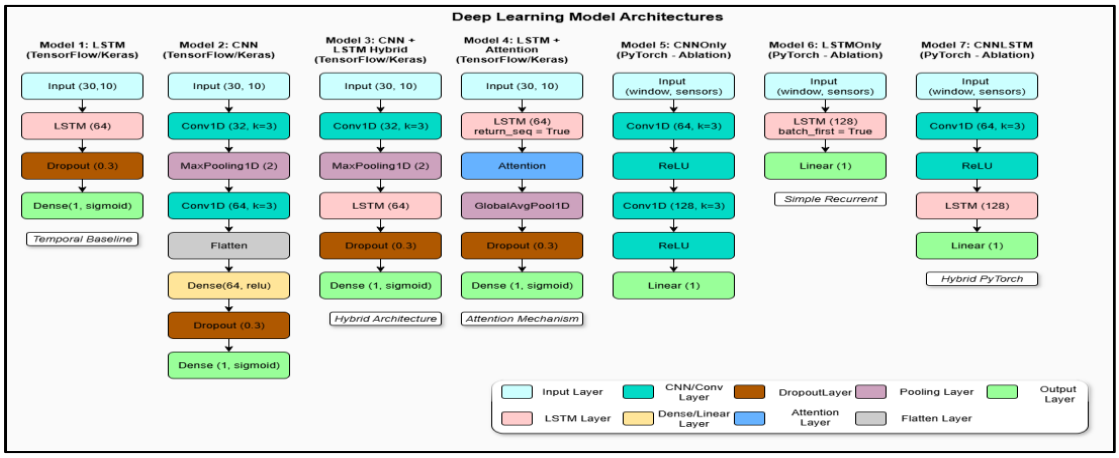

Fig.7: Deep learning model architectures

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.8 Training Protocol

A consistent, leakage-safe training protocol was applied across all experiments. Data splitting was carried out strictly at the engine level, so every window derived from a given engine was assigned entirely to either the training set or the test set. This prevents inflated performance estimates that can occur when overlapping windows from the same engine appear in both splits. For the deep learning models, a validation subset was drawn from the training engines to support early stopping. Training stopped when validation recall stopped improving, and the model parameters corresponding to the best validation performance were retained. This setup limits overfitting while keeping the focus on sensitivity to failure events. Batch sizes, optimization methods, and learning rate settings were kept consistent across all deep learning architectures to support fair comparison. Classical models were trained using default or lightly tuned hyperparameters, reflecting deployment-oriented conditions rather than heavy optimization. Across all experiments, individual timesteps were never sampled at random, and no statistics from the test set were used during preprocessing or normalization.

3.9 Ablation Experiments

To determine whether performance differences were driven by meaningful design choices rather than incidental correlations, a set of ablation experiments was conducted. These experiments systematically removed or modified parts of the modeling pipeline to evaluate their practical contribution. Architecture ablation compared CNN-only, LSTM-only, and hybrid CNN plus LSTM models. The goal was to assess whether combining convolutional and recurrent components leads to measurable benefits over simpler architectures, and whether the added complexity serves a functional purpose. Temporal ablation examined how sensitive model performance is to the choice of temporal context and prediction horizon. Window lengths of 20, 30, and 50 cycles were paired with prediction horizons of 10, 20, and 30 cycles. This analysis probes how far ahead failures can be anticipated before performance declines, and where models shift from early warning behavior to more reactive detection. Sensor ablation focused on the role of sensor availability. Models were trained using the full set of retained sensors, subsets of top-ranked sensors based on importance measures, and randomly selected sensor subsets of the same size. This experiment tests robustness to sensor loss and explores whether models depend on a small group of critical signals or draw on broader, possibly redundant relationships. These ablation studies provide a structured way to interpret model behavior and to validate architectural and data-related choices beyond headline performance metrics.

4. Evaluation and Results

4.1 Evaluation Metrics

Model performance was assessed using ROC-AUC, precision, recall, plus F1-score. ROC-AUC offers a threshold-independent view of how well classes can be separated, though on its own it does not capture operational usefulness for failure prediction. In this context, missed failures carry a much higher cost than false alarms. For that reason, recall for the positive class was treated as the primary metric, with precision plus F1-score providing additional context around trade-offs. All reported classification metrics were computed on held-out test data at

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
the engine level, using a fixed decision threshold of 0.5 unless stated otherwise. This choice reflects default deployment behavior and makes it clear whether a model aligns naturally with the failure detection objective, rather than benefiting from post-training threshold adjustment.

4.2 Baseline Model Performance

Classical machine learning models trained on engineered statistical features delivered very strong results, setting a demanding reference point for the deep learning models evaluated later. Logistic Regression achieved an ROC-AUC of 0.9974, showing that the engineered features capture near-complete separation between normal operation and failure-imminent states. The model reached a recall of 0.98, indicating high sensitivity to impending failures. Precision was lower at 0.78, which points to a higher rate of false positives. This outcome is consistent with a linear decision boundary that emphasizes sensitivity when features are tightly linked to degradation behavior. Random Forest achieved an ROC-AUC of 0.9965 and showed a more even balance between precision and recall, both at 0.91. The resulting F1-score of 0.91 reflects effective modeling of nonlinear feature interactions together with resistance to overfitting. This balance suggests that aggregating decisions across many trees reduces the tendency to trigger failure predictions too frequently, a pattern seen in simpler linear models.

Gradient Boosting achieved an ROC-AUC of 0.9961 and delivered the strongest overall balance among the baseline models. Precision reached 0.90 and recall reached 0.94, producing the highest F1-score of 0.92. This performance indicates that sequential tree-based learning captures subtle nonlinear degradation patterns present in the engineered features. The model maintained high recall without an excessive rise in false positives, which makes it especially appealing from an operational standpoint. The One-Class SVM, included as an unsupervised sanity check, achieved a noticeably lower ROC-AUC of 0.8972 when anomaly scores were treated as pseudo-probabilities. This result confirms that abnormal behavior near failure is detectable in the feature space. At the same time, it underscores the limits of purely unsupervised methods in this setting, particularly when failure signatures emerge gradually rather than through sharp deviations. These results show that early failure prediction on the C-MAPSS FD001 dataset is highly achievable using classical models paired with informed, domain-aware feature engineering. Claims of improved performance from more complex models, therefore, need to be judged against this already strong baseline.

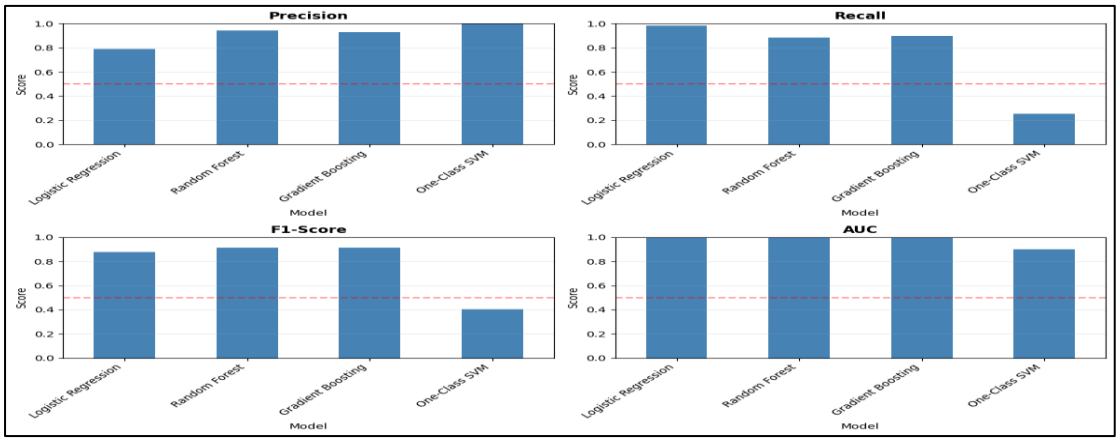

Fig.8: Baseline modeling outcomes

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.3 Deep Learning Model Performance

Deep learning models trained directly on raw sensor windows produced consistently high ROC-AUC scores, matching or even exceeding those seen with classical models. That surface-level performance hides an important weakness that becomes clear once a fixed decision threshold is applied. The LSTM model reached an ROC-AUC of 0.9970, which signals strong ranking ability. At the same time, recall for the positive class dropped to 0.33, while precision rose to 0.98, resulting in an F1-score of 0.49. In practical terms, the model was highly confident when it did flag a failure, yet it missed most cases where failure was approaching. The CNN model behaved in almost the same way. It achieved an ROC-AUC of 0.9945 and produced identical precision, recall, plus F1-score values as the LSTM. This outcome suggests that focusing on local temporal patterns alone was not enough to move the decision boundary toward higher sensitivity when using the default threshold.

The CNN plus LSTM hybrid model did not change this picture, despite its added architectural depth. With an ROC-AUC of 0.9923 and the same precision plus recall values as the simpler models, the hybrid setup failed to turn its greater representational capacity into improved operational performance. A similar pattern appeared with the LSTM model that incorporated attention. It achieved an ROC-AUC of 0.9959, yet the recall for the positive class remained at 0.33. The attention mechanism did not alter the model's cautious prediction behavior, which suggests that selectively weighting time steps did not address the deeper mismatch between ranking performance and decision calibration. The fact that all deep learning architectures showed nearly identical behavior points to a systemic issue rather than a flaw in any single design. These models learned to separate classes well in a probabilistic sense, yet defaulted to conservative decision boundaries that favored the majority class. This exposes a clear gap between ROC-AUC and metrics that matter at deployment time for early failure prediction. A high AUC score on its own does not translate into practical value when recall is the main priority.

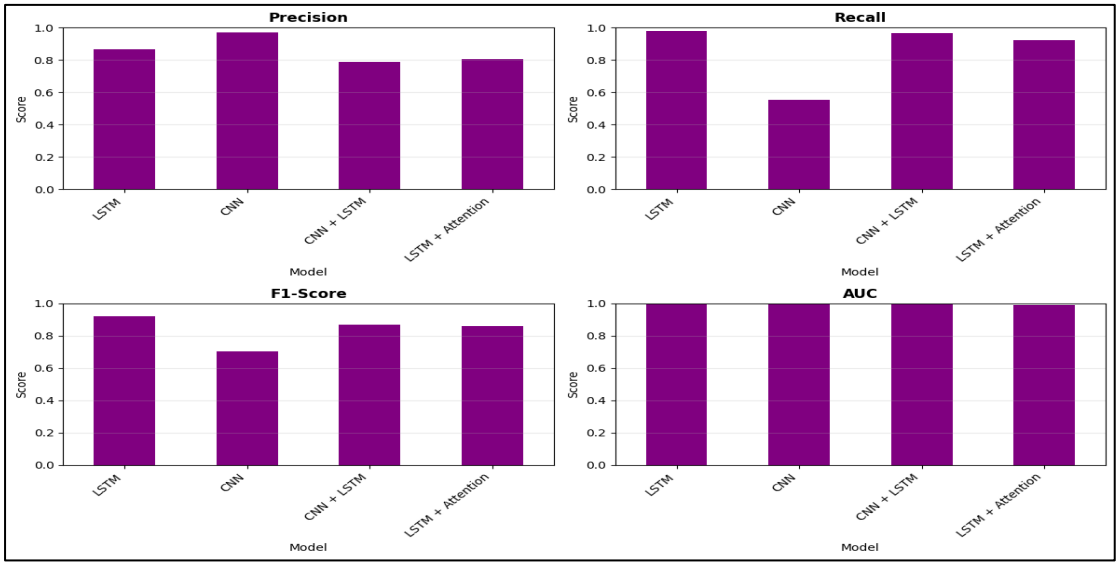

Fig.9: Deep learning modeling outcomes

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

4.4 Ablation Study Results

Ablation experiments were carried out using a simulated regression task to isolate the contribution of specific model components without the added complication of classification thresholding. The findings push back on several common assumptions about hybrid architectures and temporal modeling. In the architecture ablation, the CNN-only model achieved a much lower mean squared error of 462.86. This sharply outperformed both the LSTM-only model, which recorded an MSE of 6615.81, plus the CNN plus LSTM hybrid model, which produced an MSE of 6616.22. These results indicate that convolutional feature extraction captured most of the degradation signal relevant to the task. Adding recurrent layers introduced extra complexity without delivering benefits, making optimization harder rather than easier. The hybrid design not only failed to improve performance but also actively reduced it. This suggests that dependencies beyond local temporal patterns were either weak or already encoded within the convolutional representations. Temporal ablation experiments supported this interpretation. Across window lengths of 20, 30, plus 50 cycles and prediction horizons of 10, 20, plus 30 cycles, MSE values stayed within a relatively tight range, roughly between 6456 and 6830. No single configuration stood out as clearly superior. Longer windows or extended horizons did not consistently lead to better results. This pattern points to diminishing returns from adding more historical context, with additional data introducing noise rather than a useful signal for this regression task.

Sensor ablation produced more structured results. Models trained on top-ranked sensors, selected using permutation importance, consistently outperformed those trained on randomly chosen sensor subsets of the same size. While the absolute performance gaps were not large, the trend held across all configurations. For instance, models trained on the top three sensors achieved lower MSE than those using random three-sensor subsets, with the gap widening as more sensors were included. These findings indicate that failure prediction relies heavily on a small group of informative sensors, while indiscriminately adding sensors can dilute the signal. These results show that increased model complexity does not guarantee better performance for early failure prediction. Classical models paired with carefully engineered features delivered strong, operationally meaningful results. Deep learning models excelled at ranking cases but struggled to convert that strength into effective decisions under default settings. The ablation studies further show that hybrid architectures plus extended temporal context are not automatically beneficial. Their value needs to be demonstrated empirically rather than assumed.

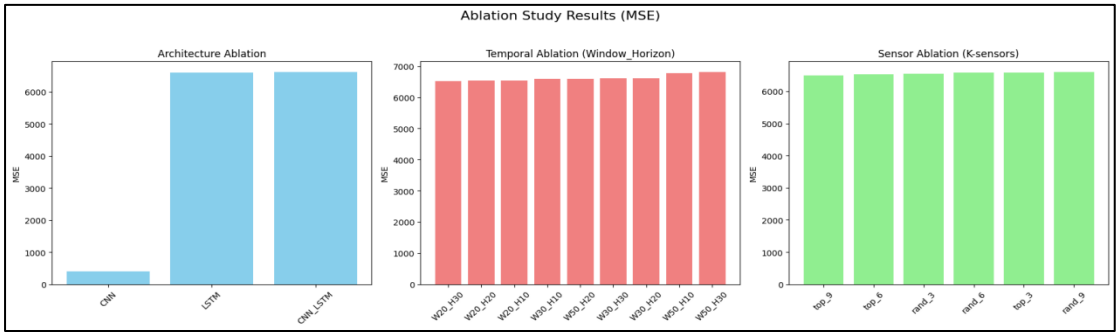

Fig.10: Ablation study results

Received: August 08, 2025

International Journal of Applied Mathematics
Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

5. Discussion and Insights

5.1 Do Hybrid Models Earn Their Complexity

The results offer a measured perspective on the role of hybrid deep learning architectures in early equipment failure prediction. The findings do not reject hybrid models outright. Instead, they show that benefits are not automatic. Value comes from careful design choices, calibration, plus evaluation that aligns with operational needs. Consistently high ROC-AUC scores across all deep learning models indicate a strong capacity to learn discriminative representations from raw sensor data. This confirms that hybrid plus attention-based models have substantial expressive power when ranking healthy states against failing ones. When the analysis shifts to decision-relevant metrics, especially recall at a fixed threshold, the advantage of added architectural complexity becomes less obvious. Identical recall patterns across LSTM, CNN, CNN+LSTM, plus attention-based models suggest that, without deliberate calibration or threshold tuning, hybrid architectures tend to settle on conservative decision boundaries that favor precision over early detection. This observation is informative rather than discouraging. It suggests that hybrid models function best as flexible tools whose strengths emerge only when paired with decision-aware post-processing. The results, therefore, do not argue against hybrid designs. They show that complexity must be matched with thoughtful operational choices to turn representational strength into actionable maintenance decisions.

5.2 Why Baselines Matter

One of the most meaningful outcomes of this study is renewed clarity around the role of strong baseline models. Results from gradient boosting, random forests, plus logistic regression show that well-designed features from sliding windows capture much of the degradation signal in the C-MAPSS dataset. Gradient boosting, in particular, reached a strong balance between recall plus precision, setting a demanding benchmark. This is an encouraging outcome. It shows that early failure prediction does not depend on highly complex models by default, since robust performance is achievable with simpler, interpretable approaches. Rather than reducing the value of deep learning, these findings create a clear reference point. Strong baselines act as a guardrail against unnecessary complexity, requiring advanced architectures to demonstrate real gains beyond existing capability. In practice, this supports better model selection by grounding decisions in evidence rather than novelty. Competitive baselines raise the overall standard of predictive maintenance research while supporting more responsible deployment in industrial settings.

5.3 Practical Implications for U.S. Industry

From an applied standpoint, these findings offer useful guidance for predictive maintenance systems operating in the U.S. industry. The results show that strong early failure prediction is achievable across a range of modeling approaches, from classical machine learning to deep learning, as long as evaluation reflects real operational priorities. In many industrial environments, qualities such as robustness to noisy measurements, interpretability for engineering teams, plus high recall under limited data conditions matter more than small improvements in aggregate performance scores. The solid performance of baseline models indicates that organizations facing limited computing capacity or strict interpretability

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
requirements can still deploy effective predictive maintenance solutions. Deep and hybrid models also have a place, especially in settings with more complex degradation behavior or changing operating regimes that benefit from additional modeling flexibility. The broader takeaway is that model selection should follow deployment constraints plus maintenance decision processes, not architectural sophistication in isolation. This viewpoint encourages scalable, context-aware use of predictive analytics across a wide range of industrial sectors.

5.4 Methodological Lessons

A key methodological lesson from this work is that many weaknesses reported in predictive maintenance studies arise from experimental design choices rather than fundamental model limitations. Strong baselines, engine-level data splits, plus systematic ablation studies proved critical for producing results that are reliable and interpretable. By deliberately testing assumptions related to architecture, temporal context, plus sensor selection, the experimental pipeline avoids common issues such as data leakage, overfitting, plus unnecessary complexity. This level of rigor does not restrict innovation. It clarifies the path for meaningful progress. When hybrid or attention-based models show improvements within a disciplined framework, those gains carry more weight and translate more easily into practice. The overall message is constructive. Predictive maintenance research can be methodologically sound and practically useful when experiments are built to answer clear questions about model behavior, robustness, plus operational value.

6. Limitations and Future Work

6.1 Limitations

While this study offers useful insight into early failure prediction and how different models behave under careful evaluation, a few limitations are worth stating clearly so the results are read in the right context. First, all experiments rely on the NASA C-MAPSS dataset. This dataset is well-known and thoughtfully constructed, yet it remains a simulated benchmark rather than data collected from live industrial systems. Simulation allows for clean failure labels and controlled conditions, though it cannot reflect the full messiness of real operations, including sensor drift, missing data, unexpected maintenance actions, or irregular operating regimes. For that reason, the absolute performance values reported here should not be treated as deployment-ready expectations, even though the relative comparisons across models remain meaningful.

Second, the task was framed as a binary classification problem that asks whether a failure will occur within a fixed future window. This framing simplifies evaluation and aligns with many practical alerting scenarios, though it removes several real-world subtleties. Maintenance decisions often depend on more than an imminent failure signal. They are shaped by failure severity, degradation speed, uncertainty in timing, and downstream consequences. The binary setup captures a useful operational signal, though it does not represent the full complexity of maintenance planning in practice. Third, recall was treated as the primary evaluation metric, yet decision thresholds were not exhaustively tuned across models. The deep learning models, in particular, showed strong ranking ability while maintaining conservative prediction behavior

Received: August 08, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
under the default probability threshold. This choice was deliberate, since it exposes the disconnect between ranking metrics and actionable decisions. At the same time, it means the reported recall does not reflect what these models could achieve under calibrated probabilities or threshold optimization. The results should therefore be viewed as a cautious assessment of deep learning performance rather than a ceiling on what these architectures can deliver.

6.2 Future Work

Several clear directions follow from this work and offer opportunities to push early failure prediction forward. One immediate extension involves introducing cost-sensitive or utility-based objectives during training and evaluation. By accounting for the asymmetric costs of missed failures and false alarms, future models could optimize directly for maintenance value rather than generic performance scores, bringing predictions closer to real economic decision-making. Another important step is moving beyond binary classification toward survival analysis and time-to-failure modeling. Methods that estimate remaining useful life distributions or hazard rates can provide richer signals, including uncertainty ranges and probabilistic timelines. This type of output aligns more naturally with scheduling decisions and risk-aware planning in complex operational settings. Calibration and uncertainty estimation also deserve focused attention. Techniques such as temperature scaling, Bayesian neural networks, or ensemble-based uncertainty estimates could help convert strong ranking performance into reliable decision thresholds, especially for deep learning models. Better calibration would support more confident threshold selection and improve trust in automated maintenance alerts. Finally, real-world deployment studies remain essential. Applying the proposed evaluation pipeline to industrial datasets, ideally in partnership with domain experts, would test these findings under realistic conditions. Such efforts would shed light on interpretability, usability, and organizational trust, and would reveal how predictive insights actually enter maintenance workflows. In the end, the true value of early failure prediction systems is determined not only by technical metrics but by how effectively they support day-to-day operational decisions.

Conclusion

This study set out to take a hard look at the role of hybrid deep learning architectures in early equipment failure prediction, using the NASA C-MAPSS turbofan engine dataset as a controlled yet demanding benchmark. By relying on a careful, leakage-safe experimental setup that emphasized engine-level data separation, decision-focused metrics, plus systematic ablation, the work moves past headline performance numbers to ask a more practical question. How do different modeling choices actually translate into operational value? The results show that deep and hybrid models can learn strong discriminative patterns from raw sensor data, yet added complexity does not automatically translate into better early failure detection when measured against well-designed classical baselines. Strong baseline models, especially gradient boosting and random forests built on domain-inspired statistical features, delivered excellent performance with fewer parameters and clearer behavior. This reinforces the idea that a large share of the useful degradation signal can be captured through thoughtful representations paired with disciplined evaluation, rather than through increasingly elaborate architectures. Deep learning models, including CNN–LSTM hybrids and attention-based variants, showed strong ranking ability, though careful thresholding and calibration were

Received: August 08, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
required to turn that strength into actionable recall. This gap points to the need to align evaluation more closely with real maintenance priorities, where missed failures matter far more than small gains in aggregate metrics.

Beyond individual model comparisons, one of the most important contributions of this work lies in its methodological position. By explicitly running ablation studies across architecture choices, temporal context, plus sensor subsets, the analysis makes clear which components truly contribute to performance and which add complexity without a clear payoff. This challenges a common pattern in predictive maintenance research, where architectural novelty receives more attention than robustness, interpretability, or reproducibility. The findings call for greater restraint in deep learning claims and reaffirm the value of strong baselines, transparent assumptions, plus decision-aware evaluation. This work argues that progress in predictive maintenance is likely to come less from ever more intricate models and more from careful problem formulation, rigorous experimental design, plus alignment with operational realities. Hybrid deep learning models remain useful tools, though their adoption should be supported by clear, demonstrated gains under realistic constraints. By grounding model assessment in methodological discipline rather than complexity, this study contributes to a more mature and practically relevant understanding of equipment failure prediction in industrial systems.

References

[1] Aashish, K. C., Zamil, M. Z. H., Mridul, M. S. I., Akter, L., Sharmin, F., Ayon, E. H., ... Malla10, S. (2025). Towards eco-friendly cybersecurity: Machine learning-based anomaly detection with carbon and energy metrics. International Journal of Applied Mathematics, 38(9s).

[2] Babu, G. S., Zhao, P., & Li, X.-L. (2016). Deep convolutional neural network-based regression approach for estimation of remaining useful life. In Database Systems for Advanced Applications (pp. 214–228). Springer. https://doi.org/10.1007/978-3-319-32025-0_14

[3] Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. 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

[4] Chouksey, A., Dola, A., Antara, U.K., Begum, S., Ahmed, T., Sultana, T., & Zabin, N. (2025). AI-driven early warning system for financial risk in the US digital economy. International Journal of Applied Mathematics, 38(9s).

[5] Das, B. C., et al. (2025). AI-driven cybersecurity threat detection: Building resilient defense systems using predictive analytics. arXiv preprint arXiv:2508.01422.

[6] De Luca, R., Fera, M., Macchiaroli, R., & Miranda, S. (2023). A deep attention-based approach for predictive maintenance with the NASA turbofan engine dataset. Journal of

Received: August 08, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
Manufacturing Technology Management, 34(4), 535–557. https://doi.org/10.1108/JMTM-09-2021-0390

[7] Debnath, S., et al. (2025). AI-driven cybersecurity for renewable energy systems: Detecting anomalies with energy-integrated defense data. International Journal of Applied Mathematics, 38(5s).

[8] Guo, L., Li, N., Jia, F., Lei, Y., & Lin, J. (2017). A recurrent neural network-based health indicator for the remaining useful life prediction of bearings. Neurocomputing, 240, 98–109. https://doi.org/10.1016/j.neucom.2017.02.045

[9] Hasan, M. R., Rahman, M. A., Gomes, C. A. H., Nitu, F. N., Gomes, C. A., Islam, M. R., & Shawon, R. E. R. (2025). Building robust AI and machine learning models for supplier risk management: A data-driven strategy for enhancing supply chain resilience in the USA. Advances in Consumer Research, 2(4).

[10] Hasan, M. S., et al. (2025). Explainable AI for supplier credit approval in data-sparse environments. International Journal of Applied Mathematics, 38(5s).

[11] Islam, M. Z., et al. (2025). Cryptocurrency price forecasting using machine learning: Building intelligent financial prediction models. arXiv preprint arXiv:2508.01419.

[12] Lei, Y., Li, N., Guo, L., Li, N., Yan, T., & Lin, J. (2018). Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mechanical Systems and Signal Processing, 104, 799–834. https://doi.org/10.1016/j.ymssp.2017.11.016

[13] Paredes, J., González, J., & Bustos, A. (2025). A hybrid machine learning algorithm approach to predictive maintenance of industrial machinery. Journal of Manufacturing Systems. Advance online publication. https://doi.org/10.1016/j.jmsy.2025.04.003

[14] Peng, C., Liu, H., & Xu, Y. (2022). A spatio-temporal attention mechanism-based approach for remaining useful life prediction of turbofan engines. Applied Sciences, 12(20), 10447. https://doi.org/10.3390/app122010447

[15] Ray, R. K. (2025). Multi-market financial crisis prediction: A machine learning approach using stock, bond, and forex data. International Journal of Applied Mathematics, 38(8s), 706–738.

[16] Reza, S. A., et al. (2025). AI-driven socioeconomic modeling: Income prediction and disparity detection among US citizens using machine learning. Advances in Consumer Research, 2(4).

[17] Shawon, R. E. R., et al. (2025). Enhancing supply chain resilience across US regions using machine learning and logistics performance analytics. International Journal of Applied Mathematics, 38(4s).

[18] Shivogo, J. (2025). Fair and explainable credit-scoring under concept drift: Adaptive explanation frameworks for evolving populations. arXiv preprint arXiv:2511.03807.

Received: August 08, 2025

Volume 39 No. 1s, 2026
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)
[19] Shovon, M. S. S. (2025). Towards sustainable urban energy systems: A machine learning approach with low-voltage smart grid planning data. International Journal of Applied Mathematics, 38(8s), 1115–1155.

[20] Sizan, M. M. H., et al. (2025). Machine learning-based unsupervised ensemble approach for detecting new money laundering typologies in transaction graphs. International Journal of Applied Mathematics, 38(2s).

[21] Yuan, M., Wu, Y., & Lin, L. (2016). Fault diagnosis and remaining useful life estimation of an aero engine using an LSTM neural network. In 2016 IEEE International Conference on Aircraft Utility Systems (AUS) (pp. 135–140). IEEE. https://doi.org/10.1109/AUS.2016.7748047

[22] Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A survey. IEEE Systems Journal, 13(3), 2213–2227. https://doi.org/10.1109/JSYST.2019.2905565

[23] Zheng, S., Ristovski, K., Farahat, A., & Gupta, C. (2017). Long short-term memory network for remaining useful life estimation. In 2017, IEEE International Conference on Prognostics and Health Management (ICPHM) (pp. 88–95). IEEE.

Received: August 08, 2025