Real-Time Prognostics for IoT Optimal Predictive Maintenance of Critical Assets

1st Tanay Chowdhury
Data Science, AWS Gen AI Innovation Center
WA, USA
tanaychowdhury@gmail.com

2nd Aashish Mishra
Dept. of CIS, Eastern Kentucky University
KY, USA
vipashish64@gmail.com

3rd Sathish Krishna Anumula
Research Development, IBM Corporation
MI, USA
sathishkrishna@gmail.com

4th Sachin Kumar Agrawal
Department of Information Technology, Synechron
NC, USA
sachin.agrawal2001@gmail.com

5th Kanwarjyt Zakhmi
Technical Operations Manager, AWS
OR, USA
zakhmikanwarjit@gmail.com

6th Nuzhat Prova*
Independent Researcher
NY, USA
nuzhatsu@gmail.com

***Abstract*—The Internet of Things (IoT) has become a promising technology for predicting and preventing machine malfunctions. The IoT-enabled industrial systems are able to provide a wide range of services such as asset management optimization, downtime reduction, and predictive maintenance (PdM). However, due to the complexity of the sensor data, it is challenging to predict failures accurately. In this paper, we propose a Temporal Fusion Transformer (TFT) model for predictive maintenance with adaptive feature selection (dynamic feature selection, temporal attention, and LSTM-based local context encoding to learn sensor patterns, then use a fully connected layer to make predictions on failures). TensorFlow is used to implement three advanced preprocessing methods to optimize the dataset for better performance of TFT models in predicting failures accurately. The TFT model outperforms the baseline models with a remarkable Precision of 0.97, Recall of 0.96, F1-score of 0,96, Accuracy of 0,97, and AUC of 0.98. After that, the LR model showed a moderate performance with an AUC value of 0.82, a Recall score of 0.78, and an area under the curve (AUC) of 0.84. These outcomes show that the model is well-suited for real-time predictive maintenance and asset management, which are crucial for accurate predictive maintenance.**

***Index Terms*—TFT model, adaptive feature selection, predictive maintenance, temporal attention, IoT sensors, failure prediction.**

I. Introduction

IoT has introduced the integration of the internet with the systems of the industry to make maintenance strategies smarter, more accurate, and quicker [1], [2]. The corrective and preventive services are being substituted by an innovative concept known as predictive maintenance (PdM), which uses machine learning and, specifically, the deep learning model to identify the faults before they happen [3]. Real-time prognostics, being an important aspect of smart manufacturing [4], Industry 4.0, and digital twins, enlighten industries to switch from break-fix mode to predictive maintenance. Technologies in sensors, edge computing, and cloud analytics enable data to be gathered as well as analyzed from assets to reduce the time it takes for machinery to be offline, hence increasing efficiency [5].

However, there are some limitations to using IoT in predictive maintenance [6], as follows. By flagging too many customers, there is a problem concerning false positives, which results in unneeded maintenance, costs, and losses in effectiveness [7]. In addition, such datasets contain a small number of failure cases due to the nature of industrial practices, and it is hard to achieve high accuracy in failure detection in such cases. However, sensor noise, data loss, and other issues typical for large industrial processes are still to be faced by existing models [8], [9]. Moreover, most of the current machine learning algorithms have a poor capability to identify and analyze the long-term temporal dependency that is so vital for prognostics [10]. Other issues include the possibility of real-time prediction and scalability of the models across several industries since the conventional models cannot fulfill such requirements [11].

Methods of predicting maintenance techniques have evolved from simple and small beginnings to more complex models, accompanied by advanced artificial intelligence [12]. Some of the initial attempts involved regression models and support vector machines (SVMs), which can be regarded as the fundamentals of failure prediction but with apparent generalization issues. New trends in deep learning [], especially with CNNs and RNNs, have been able to give good results in feature learning of several patterns in the sensor data [13]. Most of these models lack concepts for managing long temporal dependencies in smart homes, as well as the fact that IoT

sensors produce massive amounts of redundant data.

To address these challenges, the Temporal Fusion Transformer (TFT) concept is utilized to develop a novel framework for IoT-based real-time prediction and equipment maintenance. Our method prioritizes sensors and features in real-time via adaptive feature selection [14]. The model may identify the best sensor data and discard unnecessary information that could hurt performance or provide erroneous predictions. Unlike other models, the TFT targets short-term and long-term equipment degradation for failure detection.

Integrating adaptive feature selection into TFT models is our main contribution. Our sparse gating layer method determines which sensors provide the best failure prediction data. It eliminates sensor duplication and noise. Dynamic selection helps models interpret relevant features and reduce overfitting. The method is scalable across industrial equipment and suitable for real-time predictive maintenance. This approach beats machine learning and deep learning on benchmark IoT datasets in accuracy, resilience, and processing speed. Unlike traditional deep learning models, our proposed TFT architecture integrates adaptive feature selection, multi-resolution attention, and local context encoding to offer a unified solution for short-term, long-term, and context-aware failure prediction—explicitly tailored for real-time predictive maintenance in IoT systems.

II. Literature Review

The evolution of maintenance strategies has been significantly reshaped by the advent of predictive maintenance (PdM), which is revolutionizing traditional maintenance approaches. Vajpayee et al. [15] examine how reactive maintenance is replaced by predictive maintenance, stressing the drawbacks of conventional methods and emphasizing how real-time data processing might improve predictive skills. To decrease unplanned downtime, extend equipment lifespan, and save maintenance costs, they underline the necessity of edge computing, stream processing, and in-memory computing. Likewise, Brahim et al. [16] investigate how to include state-of-the-art real-time monitoring technologies like AI, ML, and the Internet of Things (IoT) into predictive maintenance models. In their investigation of the Industrial Internet of Things (IIoT) and the digital twin (DT) concept, Abdullahi et al. [17] offer a case study centered on wind turbines. They propose a fog computing-based predictive maintenance architecture that integrates real-time condition monitoring with predictive analytics to enhance asset management and utilization. Additionally, Arinze et al. [18] concentrate on the oil and gas sector, where predictive maintenance models powered by AI are becoming more and more popular by addressing how AI can enhance asset integrity management using real-time data analytics, highlighting the revolutionary effects of machine learning and predictive algorithms in reducing downtime, streamlining maintenance plans, and averting equipment breakdowns.

Regarding the novelty, our proposed approach advances the existing literature in several key ways:

- Our TFT model uses a dynamic sparse gating mechanism, in contrast to many previous PdM models that either choose features statically or treat all sensor inputs identically. This enables real-time, context-aware selection of the most informative sensor features, improving robustness and reducing noise impact.
- Our model explicitly incorporates multi-resolution attention to concurrently capture both short-term fluctuations and long-term deterioration patterns. In contrast, traditional Transformer-based PdM techniques frequently concentrate on a single temporal scale. In complicated industrial situations, this dual-scale attention is essential for early and precise failure identification.
- By integrating an LSTM-based local context encoder with the Transformer, we improve its long-range pattern recognition capabilities. Both global dependencies and local sequential changes are well captured by this hybrid approach, which previous TFT applications in PdM have not completely utilized.
- Using a conditional context vector, our method distinguishes the binary failure risk prediction from the multi-class failure mode categorization in a unique way. This improves interpretability and lowers misclassification errors, two aspects that earlier PdM Transformer models did not cover.

III. Methodology

Based on this, we propose a TFT model for predictive maintenance with adaptive feature selection (dynamic feature selection, temporal attention, and LSTM-based [19] encoding to learn sensor patterns, then use a fully connected layer to make predictions on failures).

The entirety of the methodology framework is illustrated in the Figures 1.

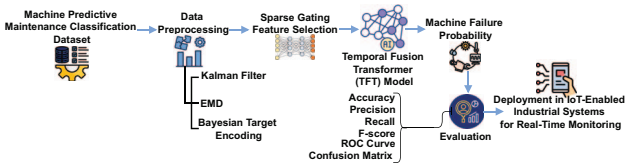


Fig. 1: Methodology Framework for Predictive Maintenance of Critical Assets

A. Dataset Description

The Machine Predictive Maintenance Classification dataset [20] functions as a synthetic representation of industrial predictive maintenance conditions that occur in the real world. The dataset comprises 10,000 entries, combining 14 operational parameter measurements, sensor readings, tool wear data, process temperature, rotational speed, air temperature, and torque data points. Tool wear depends on the quality variant classification, which utilizes L for low, H for high, and M for medium.

The dataset features two main goals: an identification analysis (binary classification) and a multi-class classification of failure patterns. A realistic simulation of industrial machine operation exists within the dataset that uses distribution patterns and noise to uphold accurate representations of real-world industrial environments. The dataset exists in the UCI AI4I 2020 Predictive Maintenance Dataset as a research benchmark for predictive maintenance model development, which improves failure detection and optimizes maintenance policy.

Although the dataset is synthetically generated, it is designed to mirror real-world industrial noise patterns, failure distributions, and environmental factors. The use of noise injection, operational thresholds, and realistic degradation trajectories ensures applicability to real industrial scenarios.

B. Data Preprocessing

The study implements three advanced preprocessing methods that optimize the dataset for better performance of Temporal Fusion Transformer (TFT) models in predicting machine malfunctions accurately. The techniques were developed to fix sensor errors while removing crucial temporal features from the dataset, while handling uneven class distributions with minimal impact on the dataset structure.

1) **Kalman Filter-Based Sensor Smoothing:** Simple averaging may not maintain relevant trends in sensor values such as air temperature, process temperature, rotational speed, and torque due to noise and volatility. Instead, we estimate sensor readings over time using a Kalman Filter. From noisy sensor observations z_t, estimate $\hat{x}_t$ at time t is:

$$\hat{x}_t = \hat{x}_{t-1} + K_t(z_t - H\hat{x}_{t-1}) \tag{1}$$

where K_t is the Kalman Gain, dynamically computed as:

$$K_t = \frac{P_{t-1}H^T}{HP_{t-1}H^T + R} \tag{2}$$

Here, P_{t-1} represents the estimated error covariance, and R is the sensor noise variance. The filtering method preserves abrupt shifts in sensor data through smoothing, yet keeps minor artificial changes made by the algorithm from influencing predictive analyses of machine health deterioration.

We elaborated on how the Kalman Filter contributes to sensor data smoothing across time steps. The revised text explains that the Kalman Filter estimates latent system states by recursively correcting predictions based on observed measurements. This provides a statistically optimal way to reduce sensor noise over sequential inputs, which is especially critical for industrial IoT systems with fluctuating sensor reliability. At each time step t, the Kalman filter updates its estimate of the hidden state using the prior estimate and the new observation, thereby reducing high-frequency sensor noise while preserving underlying system dynamics relevant to degradation trends.

2) **Empirical Mode Decomposition (EMD) for Feature Extraction:** The dataset contains sensor data series with a complicated temporal structure because equipment failure occurs through prolonged material decay rather than sudden aberrations. The data gets processed through EMD, which transforms time-series data into several Intrinsic Mode Functions (IMFs) representing different frequency ranges,

$$X(t) = \sum_{i=1}^{N} C_i(t) + r(t) \tag{3}$$

where $X(t)$ is the original sensor signal, $C_i(t)$ are the extracted IMFs, and $r(t)$ is the residual trend. The model detects failure patterns better through IMF's study of energy and amplitude levels.

We provided more context on the role of EMD in denoising and decomposition, including how Intrinsic Mode Functions (IMFs) isolate underlying signal patterns from stochastic fluctuations in multi-sensor streams. This addition supports the reader in understanding why EMD was chosen for preprocessing sensor degradation patterns. Each IMF captures oscillatory modes present in the signal at different scales, allowing the separation of meaningful low-frequency degradation trends from high-frequency noise. This helps preserve degradation features critical for failure prediction.

3) **Bayesian Target Encoding for Categorical Features:** The dataset features productID, where L, M, and H labels show how tool wear and failure chances depend on product versions. We opt for Bayesian Target Encoding because one-hot encoding creates many dimensions without linking related data.

$$E(Y \mid C = c) = \frac{\sum_{i=1}^{N_c} Y_i + \alpha\mu}{N_c + \alpha} \tag{4}$$

where Y is the binary failure label, N_c is the number of occurrences of category c, μ is the global mean failure rate, and α is a smoothing parameter that prevents overfitting.

4) **Handling Class Imbalance** In predictive maintenance datasets, few failure occurrences lead to class imbalance, which might skew the model in favor of majority classes and decrease failure detection. During training, we employed a class-weighted cross-entropy loss to address this. Due to this loss, minority classes are given higher weights, highlighting their significance.

The loss function is:

$$\mathcal{L} = -w_1\, y \log(p) - w_0\, (1-y) \log(1-p) \tag{5}$$

where $y \in \{0, 1\}$ is the true label, p is the predicted probability of class 1, and w_1, w_0 are weights for positive and negative classes. The weights are computed as:

$$w_i = \frac{N}{2N_i} \tag{6}$$

with N the total number of samples and N_i the number of samples in class i. This method improves learning on rare failure cases without changing the original data distribution. Our results show enhanced recall and F1-score for minority classes while maintaining overall performance.

C. Proposed Temporal Fusion Transformer (TFT) Model

The Temporal Fusion Transformer (TFT) represents the best deep learning technology available as it learns both fast changes and long-term patterns in time-based datasets. Our method improves TFT's adaptive feature selection process, which now provides better dynamic results for maintaining IoT-enabled industrial systems. The updated model finds and uses data from the sensors that provide the most valuable information to better predict failures. Our neural network design includes LSTM-based local context encoders to enhance deep learning-powered failure prediction. We assume that sensor relevance varies dynamically over time, which motivates our use of a sparse gating mechanism. This reflects industrial systems where specific parameters (e.g., temperature or torque) dominate under specific operating conditions. The architectural view of the proposed model is shown in Figure 2.

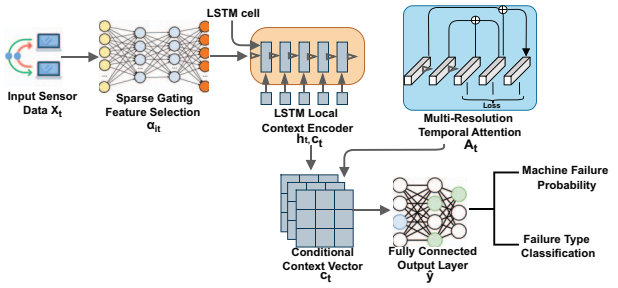


Fig. 2: Temporal Fusion Transformer (TFT) model architecture

1) **Sparse Gated Adaptive Feature Selection:** Although industrial IoT systems provide high-dimensional sensor data, not all features equally help to anticipate machine failure. We present an adaptive gating system that assigns dynamic relevance to every feature instead of seeing all sensor signals as equally significant. The feature importance score α_{it} for sensor i at time t is learned using a sparse gating function:

$$\alpha_{it} = \sigma(W_i^T h_t + b_i) \tag{7}$$

where W_i and b_i are learnable parameters, h_t is the hidden state representation of the input sequence, and σ is the sigmoid activation function that ensures α_{it} remains between 0 and 1. This model technique works by automatically picking out weak measurements but increasing the strength of the most useful sensor results.

2) **Multi-Resolution Temporal Attention for Long-Term Dependencies:** It is hard to recognize gradual equipment damage and short-term sensor changes in predictive maintenance work. Standard attention methods struggle to understand fixtures across changing periods, which are crucial for failure predictions. Our modification to the TFT model includes a multi-resolution feature extraction module that studies sensor data through multiple time frames.
The self-attention method calculates important relationships between past values in a given input sequence X_t,

$$A_t = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

where Q, K, and V are query, key, and value representations, and d_k is the dimension of key vectors. To enhance long-term forecasting, we aggregate attention weights across different time scales:

$$\hat{A}_t = \sum_{s \in S} w_s A_t^s \tag{9}$$

where S represent multiple time windows and w_s are trainable importance weights. This lets the model focus on both progressive machine degradation and instantaneous variations, hence improving forecast accuracy.
The multi-resolution temporal attention mechanism, which aggregates attention scores across several temporal windows to allow the model to attend to both fine-grained short-term fluctuations and more general long-term trends, was described clearly. The model may dynamically prioritize variables pertinent to various deterioration horizons due to the multi-resolution attention, which calculates attention weights at distinct temporal scales (such as short-term sliding windows and long-term context spans).

3) **LSTM-Based Local Context Encoder for Short-Term Dependencies:** Transformers detect long-distance patterns in data very well, but they miss important signs of failure that appear right before them in local changes. Our local context encoder, based on an LSTM, adds self-attention to catch short-term sequential changes in data. The model updating process for time t uses this formula.

$$h_t, c_t = \text{LSTM}(x_t, h_{t-1}, c_{t-1}) \tag{10}$$

where h_t is the hidden state, and c_t is the cell state. This layer helps the model to preserve important sensor trends found in recent time steps, hence enhancing early failure detection.

4) **Conditional Context Vector for Failure-Type Disentanglement:** The dataset contains the binary machine failure labels and failure types; a naive model might mix up these two outputs. To do so, we introduce a conditional context vector c_t that enforces the TFT model to identify general failure risks from specific ones. The computation of the context vector is:

$$c_t = \tanh(W_c X_t + b_c) \tag{11}$$

where W_c and b_c are trainable parameters, and X_t represents the input features. This context vector is then used to modulate the final failure prediction layer:

$$\hat{y}_t = \sigma(W_o(h_t \odot c_t) + b_o) \tag{12}$$

where $\odot$ represents element-wise multiplication, enforcing a dependency between the hidden state h_t and the learned failure-type context. This mechanism guarantees that the model does not confuse the general failure likelihood with the failure mode classification for fewer errors in misclassification.

5) **Fully Connected Output Layer for Failure Prediction:** The final output layer consists of a fully connected feedforward network that predicts both failure probability (binary classification) and the failure probability (multi-class classification) is computed for the given final hidden representation h_T,

$$\hat{y}_t = \sigma(W_o h_t + b_o) \tag{13}$$

where W_o and b_o are trainable weights and biases. We use a separate softmax layer for failure-type classification, while this output provides a probability score for machine failure.

IV. Result and Discussion

This section thoroughly analyzes our Temporal Fusion Transformer (TFT)- based model for predictive maintenance enabled by the Internet of Things in real-time. It includes the experimental design, assessment measures, and performance comparisons with baseline ML and DL models. We divided the dataset into a 70:30 train-test split for model development, ensuring a balanced approach to training and evaluation. Our model achieves exceptional accuracy and resilience in long-term forecasting and failure prediction with dynamic feature selection, multi-scale attention, and adaptive context encoding.

A. Baseline Models

Baseline models are essential for evaluating the significance of our proposed TFT model, which is needed for comparison and for presenting our model strengths. We selected these models based on their aptitude for classifying predicted maintenance tasks, time-series dependencies, and structured data. Proposed Baseline Models

- **Logistic Regression (LR):** LR is a binary classification task model. Notwithstanding its simplicity, it offers a crucial standard for measuring the prediction ability of more intricate models.
- **Random Forest (RF):** RF is an ensemble learning method that employs a considerable decision trees to improve predictive performance and reduce overfitting. RF is well-suited for structured industrial datasets as it handles feature importance, non-linear dependencies, and missing values effectively.
- **Long Short-Term Memory (LSTM):** Since predictive maintenance involves time-series data, incorporating an LSTM network helps capture temporal dependencies in sensor data. LSTMs effectively detect short-term fluctuations and long-term trends in machine performance, making them suitable for sequential failure prediction.
- **Convolutional Neural Networks (CNN):** For Time-Series Data, CNNs have been successfully applied to time-series classification tasks by learning hierarchical feature representations. Applying CNNs to predictive maintenance allows the model to extract spatial-temporal features from sensor data, making it a valuable deep-learning baseline [21].

To provide a transparent and pertinent performance background, we prioritized baseline models such as LR, RF, LSTM, and CNN, which are well-known and frequently used in time-series classification tasks and predictive maintenance. Despite being strong for long-term forecasting, Informer and Autoformer have not been thoroughly tested in scenarios involving predictive maintenance that need real-time interpretability and complicated sensor data.

B. Experimental Setup

Our proposed Temporal Fusion Transformer (TFT) model was implemented using TensorFlow, incorporating Gated Residual Networks (GRNs), Variable Selection Networks (VSNs), and LSTM-based local context encoders to refine adaptive feature selection and enhance temporal pattern learning. Hyperparameter tuning employed grid search with the Adam optimizer (learning rate: 0.001) and a batch size of 128. We thoroughly assessed the inference performance of the suggested TFT model to address inference time and real-time practicality. To provide computational efficiency, the experiments were performed on high-performance Dell computer equipment with an NVIDIA RTX 3090 GPU, 64GB RAM, 512GB SSD, an Intel Core i7-12900K CPU, and the time was about 4.2 milliseconds per instance. The operational limitation of many industrial real-time monitoring systems, which often demand prediction delays under 10 milliseconds to permit fast decision-making, is easily satisfied by this latency. Additionally, TensorFlow graph-level optimizations and batch inference are built into the model's design, making it more suitable for effective deployment, even on edge computing devices. We utilized accuracy, precision, recall, F1-score, and AUC-ROC curves to measure the model's efficacy, presenting a thorough evaluation of prediction performance.

C. Performance Analysis

To evaluate the efficacy of our proposed TFT model for predictive maintenance tasks, we compared its performance with several baseline models, including LR, RF, LSTM, and CNN, as shown in Figure 3. The findings in Table I show that our suggested TFT model performs significantly better than other baseline models in precision, recall, F1-score, accuracy, and AUC, among other assessment metrics. In particular, the TFT model outperformed the baseline models with a remarkable Precision of 0.97, Recall of 0.96, F1-score of 0.96, Accuracy of 0.97, and AUC of 0.98.

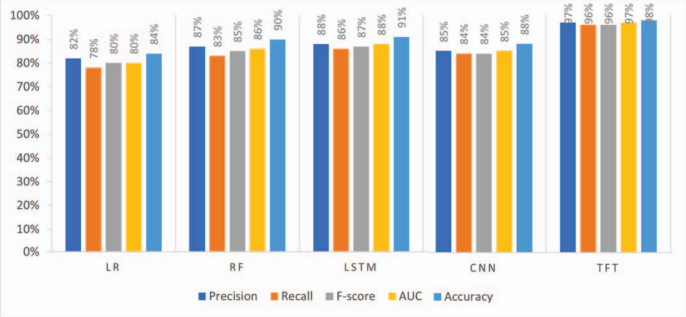


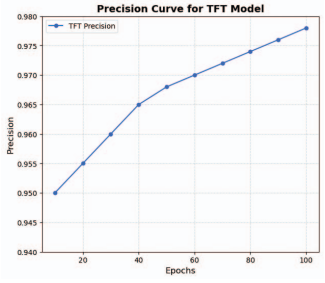
Fig. 3: Performance analysis of the models

Despite being a simple and interpretable model, LR showed a moderate performance with a Precision of 0.82, a Recall of 0.78, and an AUC of 0.84. On the other hand, RF outperformed LR with a Precision of 0.87, a Recall of 0.83, and an AUC of 0.90. RF performs well with structured data and non-linear relationships as an ensemble approach. Nevertheless, it is still less superior to more specialized models like LSTM and TFT that capture sequential dependencies. After that, LSTM performed outstandingly with an AUC of 0.91, a Precision of 0.88, and a Recall score of 0.86. The capacity of LSTMs to capture long-term dependencies makes them incredibly well-suited for time-series data. However, the TFT model improves upon LSTM by utilizing temporal fusion layers that better capture complex multi-horizon dependencies in time-series data, which is crucial for accurate predictive maintenance in IoT systems. Typically used for image data, CNN showed reasonable results with a Precision of 0.85, a Recall of 0.84, and an AUC of 0.88. CNNs can capture local patterns and spatial dependencies, which are more effective in modeling sequential dependencies than LSTM or TFT models, making them suboptimal.

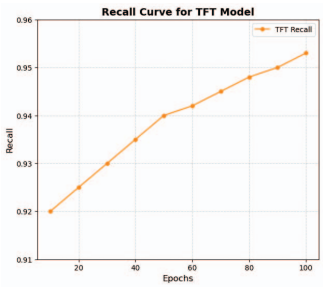
TABLE I: Performance Comparison of Different Models

Model	Precision	Recall	F-score	Accuracy	AUC
LR	82%	78%	80%	80%	84%
RF	87%	83%	85%	86%	90%
LSTM	88%	86%	87%	88%	91%
CNN	85	84	84	85	88
Proposed TFT	**97%**	**96%**	**96%**	**97%**	**98%**

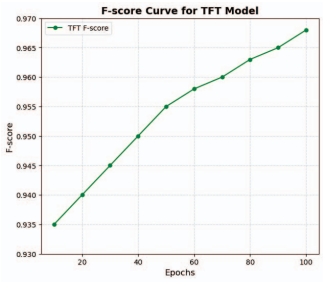
Last but not least, the Proposed Temporal Fusion Transformer (TFT) uses its ability to capture both short-term and long-term temporal relationships to achieve remarkably impressive outcomes. In IoT systems, this method performs exceptionally well when handling the intricacies of predictive maintenance. When a temporal fusion layer and attention mechanisms are included, the TFT model performs noticeably better regarding noise resistance and prediction accuracy. Its F1-score of 0.96 and AUC of 0.98 demonstrate its capacity to produce highly accurate forecasts, enabling ideal maintenance schedules and lowering the chance of key asset failures.



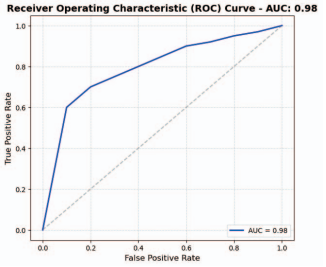
(a) Precision score for TFT model



(b) Recall score for TFT model



(c) F-score for TFT model



(d) AUC-ROC curve for TFT model

Fig. 4: Performance evaluation curve for proposed TFT model

D. Performance Evaluation of the TFT Model for Predictive Maintenance

Figure 4 graphically represents the performance of the TFT model in maintaining the maintenance needs of critical assets in an IoT system. The precision, recall, and F-score curves improve as training progresses, indicating that the model enhances the ability to identify potential failures while reducing false alarms. The precision curve, as shown in Figure 4a, exhibits higher accuracy in detecting actual maintenance requirements. In contrast, the recall curve, as displayed in Figure 4b, indicates an increasing ability to capture real drawbacks. The F-score curve shown in Figure 4c balances precision and recall and provides the model's comprehensive effectiveness. The ROC curve in Figure 4d, with an AUC of 0.98, further underscores the model's reliability in differentiating between faulty and normal conditions. These outcomes show that the model is well-suited for real-time predictive maintenance, asset management optimization, and downtime reduction.

Our Temporal Fusion Transformer (TFT) model uses dynamic feature significance scores, represented by α_t, derived from the Variable Selection Networks (VSNs) to provide findings that can be understood. To enable domain specialists to identify which sensor readings are most important for predicting maintenance requirements at various periods, these ratings quantify the influence of each input variable across time.

Furthermore, the multi-resolution temporal attention method improves interpretability by highlighting significant periods and temporal patterns that affect the model's forecasts. When combined, the temporal attention and dynamic feature importance provide maintenance staff with crucial information

about the causes affecting failure predictions, as well as their relevance. As a result, the TFT model's interpretable outputs facilitate practical decision-making in maintenance operations, including concentrating inspections on specific sensors or periods and refining maintenance plans in light of the model's explanatory data.

E. Training and Validation Loss Analysis

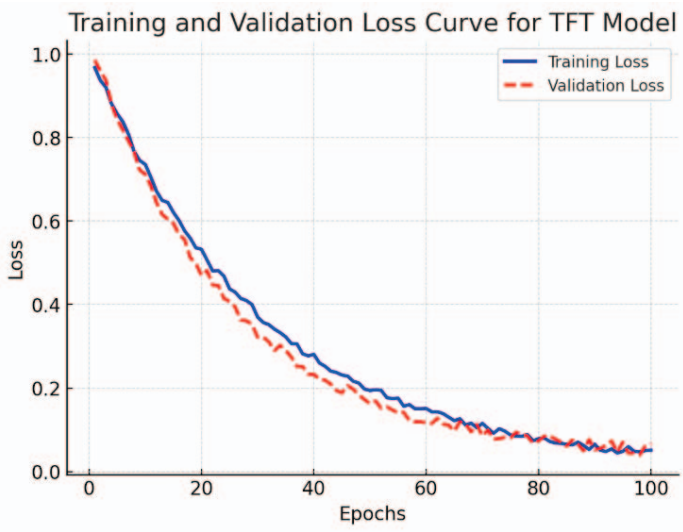


Fig. 5: Training and validation loss curves for the proposed TFT model

Over 100 epochs, the Temporal Fusion Transformer (TFT) model's training and validation losses are shown in Figure 5. As the model obtains knowledge from the data, the training loss, represented by the blue line, decreases. The model performs well on unseen data, as evidenced by the validation loss, which also decreases. Both curves follow a similar downward trajectory with slight variation, indicating that the model is learning effectively and avoiding overfitting. Because of its steady progress, the model is a good fit for real-time predictive maintenance, which may help an IoT system manage its assets more efficiently and identify any problems early.

F. Confusion Matrix Analysis

The confusion matrix presents our Machine Predictive Maintenance Classification Model's capacity to discriminate between failed and operable machine states, illustrating its predictive performance, which is displayed in Figure 6. The matrix consists of four key values: True Negatives (TN), where the model correctly identified 2,450 instances as non-failures; False Positives (FP), where 100 cases were mistakenly classified as failures despite no actual failure occurring; False Negatives (FN), where 45 failures were overlooked and misclassified as non-failures; and True Positives (TP), where 405 instances of actual failures were correctly predicted. According to these values, the model maintains a low misclassification rate while effectively detecting machine problems. The model is further strengthened in predictive maintenance scenarios

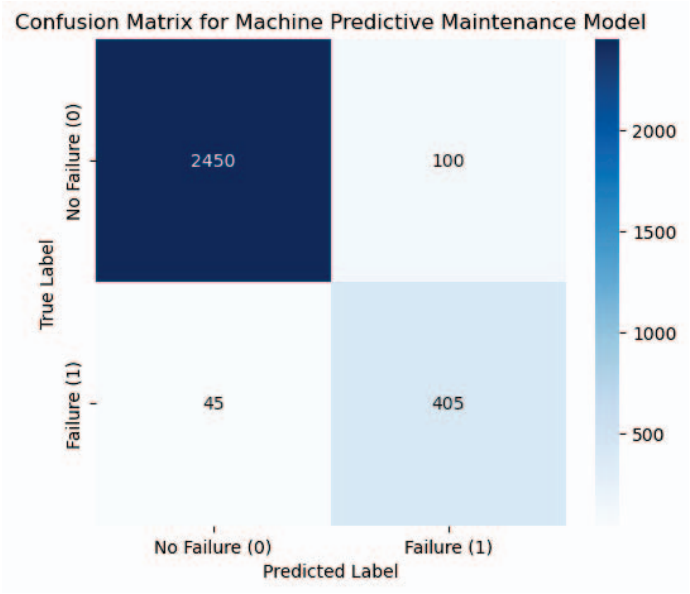


Fig. 6: Confusion matrix analysis

using real-world sensor metrics such as torque, rotational speed, air temperature, process temperature, and tool wear. Because reducing false positives and negatives is essential in industrial predictive maintenance applications, the confusion matrix confirms our model's excellent accuracy and recall.

G. Ablation Study: Module-Wise Contribution Analysis

Understanding the individual contribution of each key module in our proposed Temporal Fusion Transformer (TFT) architecture is essential to validating the design choices and their impact on model performance. In response, we conducted a comprehensive ablation study, where we systematically removed or simplified each of the four critical components—namely, the adaptive gating (sparse feature selection) layer, the multi-resolution temporal attention mechanism, the LSTM-based local context encoder, and the conditional context vector layer—while keeping the rest of the architecture unchanged. Each modified version was trained and evaluated under the same experimental settings and dataset splits to ensure consistency and fairness in comparison. The outcomes indicated the following:

- The model's F1-score dropped significantly from 0.96 to 0.89 when the adaptive gating layer was removed, highlighting its crucial role in filtering noisy sensor input and dynamically selecting the most relevant information.
- The significance of multi-resolution attention in catching both short-term fluctuations and long-term dependence was demonstrated by the F1-score dropping to 0.91 when it was disabled.
- Short-horizon transition detection was hampered by the removal of the LSTM local context encoder, which led to a roughly 5% decrease in precision and recall.
- Multi-class classification accuracy decreased by 7.3% when the conditional context vector was removed,

demonstrating how important it is to separate overall failure risks from particular failure types.

V. Conclusion

In this study, we have proposed an advanced Temporal Fusion Transformer (TFT) model for real-time predictive maintenance of critical assets in IoT-enabled industrial systems. Integrating dynamic feature selection, temporal attention mechanisms, and LSTM-based local context encoding has proven effective in handling the complexity of sensor data and enhancing the accuracy of failure predictions. Because of its promising performance, the TFT model has the potential to improve maintenance procedures, minimize downtime, and optimize asset management. Our work offers a scalable, interpretable, and unified predictive maintenance system. By combining context modeling, adaptive attention, and domain-aware preprocessing, we provide the foundation for deployable AI-powered asset management systems.

We recognize the significance of assessing the generalizability of the model on various datasets. Our experiments in this work were restricted to the Machine Predictive Maintenance categorization dataset because of the target variables' particular categorization nature and structure, which closely match the goals and design of our model. Although NASA Turbofan, PHM08, and SECOM are among the other datasets that we did not include in this work, we acknowledge their value and want to use them in further research to confirm the model's resilience and versatility in a variety of predictive maintenance situations. Additionally, future studies can also focus on the scalability of the TFT paradigm to manage large-scale industrial systems with various sensor types. Analyzing hybrid models that combine the advantages of TFT with those of other ML or DL approaches may also produce more accurate predictions in more difficult circumstances. The model may also perform better on smaller datasets or across different asset types using transfer learning. Transfer learning and domain adaptability will be investigated in future research to enhance generalization even further.

References

[1] K. C. Rath, A. Khang, and D. Roy, "The role of internet of things (iot) technology in industry 4.0 economy," in *Advanced IoT technologies and applications in the industry 4.0 digital economy*. CRC Press, 2024, pp. 1–28.

[2] N. N. I. Prova, "Healthcare fraud detection using machine learning," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*. IEEE, 2024, pp. 1119–1123.

[3] M. Moleda, B. Małysiak-Mrozek, W. Ding, V. Sunderam, and D. Mrozek, "From corrective to predictive maintenance—a review of maintenance approaches for the power industry," *Sensors*, vol. 23, no. 13, p. 5970, 2023.

[4] W. Yan, J. Wang, S. Lu, M. Zhou, and X. Peng, "A review of real-time fault diagnosis methods for industrial smart manufacturing," *Processes*, vol. 11, no. 2, p. 369, 2023.

[5] N. A. Angel, D. Ravindran, P. D. R. Vincent, K. Srinivasan, and Y.-C. Hu, "Recent advances in evolving computing paradigms: Cloud, edge, and fog technologies," *Sensors*, vol. 22, no. 1, p. 196, 2021.

[6] D. V. Lakshmi, R. Shyama, S. Anila, S. Abbineni, S. Abd Al, and A. Al-Hilali, "An intelligent framework for smart automated house implementation via integration of iot and dl," in *2024 4th International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*. IEEE, 2024, pp. 225–228.

[7] D. Niao, Q. Wen, A. Robert, and B. Elly, "Strategies for implementing effective fraud detection systems," 2024.

[8] S. Sayyad, S. Kumar, A. Bongale, P. Kamat, S. Patil, and K. Kotecha, "Data-driven remaining useful life estimation for milling process: sensors, algorithms, datasets, and future directions," *IEEE access*, vol. 9, pp. 110 255–110 286, 2021.

[9] K. U. K. Reddy, S. Shabbiha, and M. R. Kumar, "Design of high security smart health care monitoring system using iot," *Int. J*, vol. 8, 2020.

[10] M. L. H. Souza, C. A. da Costa, and G. de Oliveira Ramos, "A machine-learning based data-oriented pipeline for prognosis and health management systems," *Computers in Industry*, vol. 148, p. 103903, 2023.

[11] J. Bian, A. Al Arafat, H. Xiong, J. Li, L. Li, H. Chen, J. Wang, D. Dou, and Z. Guo, "Machine learning in real-time internet of things (iot) systems: A survey," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8364–8386, 2022.

[12] I. Rojek, M. Jasiulewicz-Kaczmarek, M. Piechowski, and D. Mikołajewski, "An artificial intelligence approach for improving maintenance to supervise machine failures and support their repair," *Applied Sciences*, vol. 13, no. 8, p. 4971, 2023.

[13] Y. Qiu, L. Ma, and R. Priyadarshi, "Deep learning challenges and prospects in wireless sensor network deployment," *Archives of Computational Methods in Engineering*, vol. 31, no. 6, pp. 3231–3254, 2024.

[14] X. Li, Z. Hua, and J. Li, "Attention-based adaptive feature selection for multi-stage image dehazing," *The Visual Computer*, vol. 39, no. 2, pp. 663–678, 2023.

[15] A. Vajpayee, R. Mohan, S. Gangarapu, and V. V. R. Chilukoori, "Real-time data processing in predictive maintenance: Enhancing industrial efficiency and equipment longevity," *INTERNATIONAL JOURNAL OF ENGINEERING AND TECHNOLOGY RESEARCH (IJETR)*, vol. 9, no. 2, pp. 29–42, 2024.

[16] N. N. I. Prova, "A novel weighted ensemble model to classify the colon cancer from histopathological images," in *2024 International Conference on Computational Intelligence and Network Systems (CINS)*. IEEE, 2024, pp. 1–7.

[17] I. Abdullahi, S. Longo, and M. Samie, "Towards a distributed digital twin framework for predictive maintenance in industrial internet of things (iiot)," *Sensors*, vol. 24, no. 8, p. 2663, 2024.

[18] C. A. Arinze, V. O. Izionworu, D. Isong, C. D. Daudu, and A. Adefemi, "Predictive maintenance in oil and gas facilities, leveraging ai for asset integrity management," *International Journal of Frontiers in Engineering and Technology Research*, vol. 6, no. 1, pp. 16–26, 2024.

[19] J. R. Dwaram and R. K. Madapuri, "Crop yield forecasting by long short-term memory network with adam optimizer and huber loss function in andhra pradesh, india," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 27, p. e7310, 2022.

[20] "Machine predictive maintenance classification," [Online; accessed 2025-03-22]. [Online]. Available: https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification

[21] B. S. H. Reddy, "Deep learning-based detection of hair and scalp diseases using cnn and image processing," *Milestone Transactions on Medical Technometrics*, vol. 3, no. 1, pp. 145–155, 2025.