A Transformer-Based Framework for Early Fault Detection in Industrial Sensor Networks

Lucas A. Verhoeven^{1,2,#}, Priya N. Raghunathan^{1,#}, Samuel K. Adeyemi^{1,2}, Hana Lindqvist¹, Marco T. Ferreira^{1,2,*}

¹Department of Intelligent Systems and Industrial Analytics, Lakemont University of Technology, Brevik, Norway; ²Laboratory for Predictive Maintenance Research, Lakemont University of Technology, Brevik, Norway

Correspondence to: Dr. Marco T. Ferreira, Department of Intelligent Systems and Industrial Analytics, Lakemont University of Technology, Brevik, Norway; Email: m.ferreira@lakemont.example.edu

ABSTRACT

Early detection of equipment faults from multivariate sensor streams remains a central challenge in industrial analytics, particularly when static machine metadata must be analyzed alongside continuously recorded signals. This study proposes a transformer-based architecture, combining a sensor embedding layer, a multi-head self-attention encoder, and a metadata fusion head, for joint fault classification across heterogeneous inputs. The proposed model was evaluated against three established baselines (support vector machine, random forest, and LightGBM) on a combined metadata and sensor-stream benchmark of 62,300 records. The proposed model achieved 96.1% accuracy and an F1-score of 0.953, outperforming the strongest baseline (LightGBM) by 4.7 percentage points, at a modest increase in inference time (11.8 ms versus 7.2 ms). Ablation analysis showed that removing the transformer encoder reduced accuracy by 6.8 points and restricting the model to single-head attention reduced accuracy by 3.1 points, indicating that both sequence-level attention and multi-head feature weighting contribute meaningfully to overall performance. These results suggest that transformer-based architectures offer a practical and accurate approach for fault detection tasks involving mixed static and sequential data.

Keywords: Fault detection · Predictive maintenance · Transformer · Self-attention · Industrial sensor networks · Deep learning

INTRODUCTION

The rapid expansion of connected industrial equipment has increased demand for analytical models that can process static machine metadata and streaming sensor signals within a single pipeline. Conventional approaches typically treat these inputs separately, applying tree-based methods to metadata and recurrent or convolutional networks to signals, which can limit their ability to capture interactions between machine characteristics and temporal behavior [2;5].

Gradient-boosted trees and random forests remain widely used in industrial settings because of their strong performance, robustness to missing values, and low computational cost [3;5]. However, these methods rely on hand-crafted summary statistics to represent time-dependent behavior, which may discard fine-grained temporal patterns that precede equipment failure. Transformer architectures have achieved strong results in language and vision tasks, but their application to mixed metadata and sensor-stream data in industrial contexts has been comparatively limited [6;8].

This study proposes a transformer-based architecture designed to jointly process static

© 2026 The Information Science Journal

metadata and multivariate sensor sequences for fault classification. The purpose of this study was to evaluate the predictive accuracy of the proposed model relative to established baselines, and to quantify the individual contribution of its components through ablation analysis.

METHODS

Dataset

A combined benchmark dataset of 62,300 records was assembled from machine metadata and associated sensor streams drawn from a publicly available industrial condition-monitoring corpus. Each record included 36 static metadata features and a corresponding 48-step sequence of 12 sensor channels. The dataset was partitioned into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve class balance across fault categories.

Model architecture

The proposed model consists of three components. A linear embedding layer first projects each time step of the sensor sequence into a 128-dimensional representation, with sinusoidal positional encoding added to preserve temporal order. The embedded sequence is passed to a four-layer transformer encoder with eight attention heads per layer, allowing the model to assign learned weights to time steps and channels most relevant to the fault target. The pooled sequence representation is then concatenated with the static metadata features and passed to a fully connected classification head.

Baseline models

Three baseline models were trained for comparison: a support vector machine (RBF kernel), a random forest (500 estimators), and LightGBM (500 estimators, maximum depth 7). All baselines were trained on the metadata features together with engineered summary statistics of each sensor channel (mean, variance, minimum, maximum, and slope), consistent with standard practice for these model classes. Hyperparameters were tuned using grid search on the validation set.

Training and evaluation

The proposed model was trained using the AdamW optimizer with an initial learning rate of 5e-4, a batch size of 256, and early stopping based on validation loss over a maximum of 120 epochs. Performance was evaluated on the held-out test set using accuracy, F1-score, and mean inference time per record. An ablation study was conducted by systematically removing the transformer encoder, restricting the encoder to single-head attention, and removing the noise-injection augmentation procedure applied during training.

RESULTS

Predictive performance

The proposed model achieved the highest performance among all evaluated approaches, with 96.1% accuracy and an F1-score of 0.953 on the held-out test set (Table 1). This represented an improvement of 4.7 percentage points in accuracy over the strongest baseline, LightGBM (91.4%), and a larger margin over random forest (87.6%) and the support vector machine (80.2%). The improvement was accompanied by an increase in mean inference time, from 7.2 ms for LightGBM to 11.8 ms for the proposed model.

Model	Input	Accuracy (%)	F1-Score	Inference Time (ms)
Support Vector Machine (baseline)	Metadata + summary statistics	80.2	0.771	3.4
Random Forest	Metadata + summary	87.6	0.852	5.9

© 2026 The Information Science Journal

Model	Input	Accuracy (%)	F1-Score	Inference Time (ms)
	statistics			
LightGBM	Metadata + summary statistics	91.4	0.899	7.2
Transformer (proposed)	Metadata + raw sensor sequence	96.1	0.953	11.8

Table 1: Predictive performance of the proposed model compared with baseline models on the held-out test set.

Ablation analysis

Ablation experiments indicated that all three components contributed to overall performance (Table 2). Removal of the transformer encoder, leaving only a feed-forward network over pooled sensor features, produced the largest decrease, reducing accuracy by 6.8 percentage points to 89.3%. Restricting the encoder to single-head attention reduced accuracy by 3.1 points to 93.0%, and removal of the noise-injection augmentation reduced accuracy by 4.2 points to 91.9%. These results suggest that sequence-level attention contributed most substantially to predictive performance, while multi-head weighting and augmentation provided smaller but consistent additional gains.

Ablation Configuration	Accuracy (%)	Δ from Full Model
Full proposed model	96.1	—
Single-head attention only	93.0	-3.1
Without transformer encoder (feed-forward only)	89.3	-6.8
Without noise-injection augmentation	91.9	-4.2

Table 2: Ablation study isolating the contribution of individual components to model accuracy.

DISCUSSION

The results indicate that a transformer-based architecture can outperform established ensemble baselines on fault classification tasks involving mixed static and sequential data. The magnitude of improvement over LightGBM, the strongest baseline evaluated, suggests that the raw sensor sequence contains predictive signal that is not fully preserved by hand-crafted summary statistics.

The ablation analysis further clarifies the contribution of each component. The large decrease observed when removing the transformer encoder indicates that sequence-level attention is a primary driver of the model's advantage, while multi-head attention appears to provide a secondary refinement by allowing different heads to specialize in distinct sensor channels or time scales. The contribution of data augmentation, though smaller, was consistent with its role in improving generalization to unseen machines.

The increase in inference time associated with the proposed model, while modest in absolute terms, may matter for deployment on resource-constrained edge devices. Organizations should weigh whether the accuracy gains justify the additional computational cost for their specific application. This study has several limitations, including reliance on a single

© 2026 The Information Science Journal

benchmark dataset and a fixed train-validation-test split; future work should evaluate the architecture across additional equipment types and assess its robustness under sensor drift and missing data.

CONCLUSION

This study presented a transformer-based architecture for fault detection involving combined static metadata and multivariate sensor streams, and demonstrated that it outperforms conventional ensemble baselines by a meaningful margin. Ablation analysis confirmed that both sequence-level attention and multi-head feature weighting contribute to the observed gains. These findings support the use of attention-based architectures for predictive maintenance tasks, while highlighting the trade-off between accuracy and inference latency that should inform deployment decisions.

DECLARATIONS

Funding: This study was supported by the Lakemont University of Technology Industrial Analytics Fund [LUT-IAF-2026-014] and the Brevik Regional Manufacturing Grant [BRMG-2207].

Conflicts of Interest: The authors declare no conflicts of interest.

Data Availability Statement: The benchmark dataset and processed model outputs supporting the findings of this study are available from the corresponding author upon reasonable request.

Author Contributions: Conceptualization, Ferreira M.T. and Verhoeven L.A.; methodology, Raghunathan P.N.; software, Adeyemi S.K.; formal analysis, Lindqvist H.; investigation, Verhoeven L.A.; writing—original draft, Verhoeven L.A.; writing—review and editing, Ferreira M.T. and Raghunathan P.N.; supervision, Ferreira M.T. All authors have read and agreed to the published version of the manuscript.

REFERENCES

1. Verhoeven L.A., Raghunathan P.N. Attention-based models for industrial condition monitoring. J Ind Data Sci. 2025;11(2):44-62.
2. Adeyemi S.K., Lindqvist H. Tree ensembles for predictive maintenance: strengths and blind spots. Ann Mach Intell. 2024;6(3):118-135.
3. Ferreira M.T., Verhoeven L.A. Hand-crafted versus learned features in sensor analytics. Appl Signal Learn. 2025;8(1):21-39.
4. Raghunathan P.N., Adeyemi S.K. Multivariate time-series fusion with static covariates. Comput Mfg Syst. 2024;15(4):203-221.
5. Lindqvist H., Ferreira M.T. Gradient boosting in factory deployments: a practitioner survey. Data Eng Rev. 2023;4(2):57-74.
6. Verhoeven L.A., Lindqvist H. Transformers beyond language: a review of sequential industrial applications. Neural Syst Appl. 2025;10(3):88-109.

© 2026 The Information Science Journal