HyLME: Language Model Embeddings with Knowledge Distillation for Robust Predictive Maintenance under Missing Sensor Data

Ju-Young Kim
Dept. of Computer Science and Engineering
Gyeongsang National University
Jinju, Republic of Korea
wndudwkd003@gnu.ac.kr

Ji-Hong Park
Dept. of Computer Science and Engineering
Gyeongsang National University
Jinju, Republic of Korea
hong_0002@gnu.ac.kr

Gun-Woo Kim*
Dept. of Computer Science and Engineering
Gyeongsang National University
Jinju, Republic of Korea
gunwoo.kim@gnu.ac.kr

***Abstract*—In the 4th Industrial Revolution, predictive maintenance is a key strategic element that predicts equipment failure or status in real-time by collecting relevant data from sensors. However, missing sensor data occurs frequently, leading to enormous opportunity costs. Therefore, there is a growing need for robust models that can maintain high accuracy and stable performance even when missing data occurs. Against this backdrop, language models have achieved remarkable success in natural language processing tasks through the innovative architecture of Transformers, and their context-based learning effect has been widely proven in other domains such as tabular data and time-series analysis. In this paper, we propose a Hybrid architecture with a machine learning teacher and a Language Model Embedding-based student (HyLME) to enable accurate and robust predictive maintenance even in the presence of missing data. HyLME is a hybrid learning approach that fuses text embeddings from language models and learns knowledge from tree-based machine learning models. Additionally, we implemented a masking scenario based on feature importance to simulate missing data conditions. In the results of comparative experiments, HyLME achieved an average accuracy of 0.83207 across all scenarios. This performance is 15.99% higher than the average accuracy of the comparison models, which indicates that the proposed architecture is effective in performing accurate and robust predictions in sensor missing data situations.**

Keywords— predictive maintenance, missing data processing, language model embeddings, hybrid deep learning and machine learning, knowledge distillation

I. INTRODUCTION

Predictive maintenance is a key strategy that maximizes productivity and efficiency by monitoring internal and external data from sensors attached to equipment in real time, predicting potential failures in advance, or accurately assessing the current state to perform maintenance at the appropriate time [1,2]. However, in some environments, outliers or missing data frequently occurs in the collected data due to various factors such as equipment lifespan limitations caused by aging and physical failures caused by external shocks. Shutting down and repairing the equipment whenever missing data occur incurs enormous opportunity costs, and situations may unavoidably arise in which sensor replacement is difficult. If these problems accumulate, they leads to a serious situation where it becomes impossible to accurately predict the state of the equipment [3,4]. Therefore, the need for a robust model that maintains high accuracy and consistent performance is emerging, even when missing data occurs within a sequence of data.

Amidst the growing demand for research in the 4th Industrial Revolution and predictive maintenance, language models are gaining prominence as a rapidly emerging form of artificial intelligence. Language models began with the innovative Transformer architecture, and they have evolved into various Large Language Models (LLMs) such as BERT [5], GPT-2 [6], and Llama 3 [7]. By understanding the meaning of words and their contextual relationships in sentences, these models have demonstrated successful results in Natural Language Processing (NLP).

These Transformer-based language models possess the ability to understand contextual relationships, namely, the relationships between data points within vast datasets, and to robustly capture complex patterns. As a result, they have expanded into domains beyond NLP and have contributed to a paradigm shift. In particular, Transformers and LLMs have been actively explored recently in traditional machine learning and deep learning tasks, such as structured tabular data and time-series analysis. Based on this background, we raise the following two research questions regarding our study

1) In sensor-missing environments for predictive maintenance, can hybrid ML/DL models maintain high accuracy despite progressive sensor failures?

2) Can the text embeddings from pre-trained language models contextually capture inter-sensor relationships and accurately project complex data patterns?

In this paper, we explore the two aforementioned research questions and propose a Hybrid architecture with a machine learning teacher and a Language Model Embedding-based student (HyLME) to perform accurate and robust predictive maintenance in environments where sensor missingness frequently occurs. The proposed architecture serializes sensor data as text and extracts embeddings using a pre-trained language model. These embeddings are then fused with the original data and processed through a ResidualMLP head (a neural network with residual connections). In this process, the model is trained through knowledge distillation technique, where a tree-based machine learning model serves as the teacher and the ResidualMLP serves as the student. This approach leverages the contextual learning capability of the language model to capture inter-sensor relationships and the robustness of a tree-based model against missing data.

The main contributions of this work are as follows:

- **HyLME:** This study demonstrates that the proposed hybrid teacher-student architecture, which combines pre-trained language model embeddings with machine learning and deep learning models, is effective in achieving accurate and robust predictive maintenance even in the presence of missing sensor data.
- **Masking Scenario:** To ensure accurate and fair evaluation under sensor missing conditions, a masking scenario was constructed based on feature importance.

* Corresponding author

- **Experimental Validation:** Through comparative experiments with ten baseline models on the MPTMS dataset, HyLME achieved 15.99% higher average accuracy, demonstrating the effectiveness of the proposed framework in handling missing sensor data.

II. Related Works

A. Predictive Maintenance and Handling Missing data

The advancement of machine learning and deep learning has played a significant role in the progress of the 4th Industrial Revolution and predictive maintenance, leading to successful outcomes. Hermawan et al. [8] proposed a CLSTM architecture that combines CNN and LSTM to predict the Remaining Useful Life (RUL) of aircraft engines, successfully integrating important features from each structure and achieving high accuracy. Furthermore, Ghadekar et al. [9] proposed a machine learning-based anomaly detection model using XGBoost and Local Outlier Factor to detect failures and abnormal states in industrial equipment, demonstrating high accuracy and introducing SHAP to enhance explainability, thereby indicating the model's applicability in real-world settings. However, in actual facilities, missing sensor data occurs frequently, and both studies have the limitation of not considering such missing data scenarios.

To address missing data in tabular form, numerous studies have continuously explored methods of data interpolation and generation. Ba-Alawi et al. [10] and Lee et al. [11] utilized Auto Encoder and Variational Auto Encoder architectures, respectively, to learn from both incomplete and complete data in order to restore the missing data. Both studies commonly pointed out that traditional missing data processing methods, such as simple mean imputation or interpolation, have the limitation of failing to reflect the temporal characteristics and interrelationships within the data. Their proposed architectures demonstrated high accuracy by effectively generating normal data.

The generation of missing data stems from the objective of accurately predicting with normal data. However, Caruso et al. [12] pointed out that conventional machine learning and deep learning models can only learn by replacing missing data in advance, and that the act of interpolating or generating missing data itself may cause information loss or bias. Caruso et al. proposed the NAIM, which is designed to learn effectively from incomplete data by disregarding missing data. NAIM treats missing data as a special token and ignores their influence within the Transformer's attention mechanism. Moreover, it utilizes a regularization technique that randomly masks a subset of features during each training epoch, thereby inducing the model to learn generalized capabilities in the presence of missing data. This approach indicates that refraining from explicitly processing missing data can, in itself, serve as a novel means of improving generalization in predictive models.

B. Language Models and Text Embeddings

In a notable study on language model embeddings, Tang et al. [13], pointed out that most regression studies utilizing LLMs have focused on decoding-based approaches that generate tokens directly as outputs, while investigations employing embedding vectors directly remain insufficient. Additionally, traditional embedding approaches exhibit a steep performance drop when applied to high-dimensional table-column inputs. To mitigate this, Tang et al. serialized table data into string, extracted text embeddings using an LLMs, and fed them into an MLP regression head. Tang et al. proved that the LLM preserves Lipschitz continuity by not excessively expanding output distances relative to input distances, and they suggested that the core property of the embedding representation is its ability to maintain performance or exhibit only a gradual decline irrespective of input feature dimensionality.

In addition, Kaur et al. [14] conducted the first study to leverage text embeddings from LLMs for time-series data. They pointed out that existing time-series analysis approaches using LLMs require fine-tuning with millions of parameters, which leads to high computational and memory costs, making them unsuitable for resource-constrained environments. They also raised questions about the effectiveness of directly applying LLMs to time-series data in terms of accuracy and efficiency. LETS-C, proposed by Kaur et al., serializes sequences into strings using digit-space tokenization, which preserves numerical continuity. It then extracts embeddings using a pretrained LLMs and trains a classifier composed of CNN and MLP structures. This approach is grounded in the capabilities of pre-trained LLMs, which, owing to their Transformer structures and context-based learning methods, can produce rich sequence representations and accurately capture the complex patterns and temporal dependencies in time-series data. Consequently, LETS-C has demonstrated remarkable results, achieving State-of-the-Art (SOTA) performance in benchmark experiments across various datasets. Additionally, through cosine similarity analysis of data samples represented as text embeddings, Kaur et al. proved that samples within the same class are located close to each other, while those from different classes are positioned farther apart. This finding suggests that text embeddings from LLMs are effective for time-series data.

While these studies have made significant contributions to handling missing data and leveraging LLM embeddings separately, our work uniquely combines these approaches through knowledge distillation for robust predictive maintenance. This integration addresses the limitations of both approaches: the domain-specific knowledge gap in LLMs and the potential information loss in traditional missing data handling methods.

III. Methodology

A. HyLME: Intergration of Models Robust to Missing Data

HyLME employs a hybrid architecture in the form of knowledge distillation [15] that learns both the text embeddings extracted from a pre-trained BERT [5] model and the logits of an XGBoost [16] model, which is a tree-based machine learning algorithm.

BERT is trained with both the Masked Language Model technique, which learns local contextual information within a sentence, and the Next Sentence Prediction technique, which learns global contextual information between sentences. This training approach makes it possible to effectively capture complex semantic relationships between words and sentences. It has a characteristic where the overall interpretation of the context is not significantly affected, even if some words in a sentence are inappropriate, such as typos. This robustness to missing or corrupted text translates well to handling missing sensor readings in our application.

This implies that even in the presence of missing data or

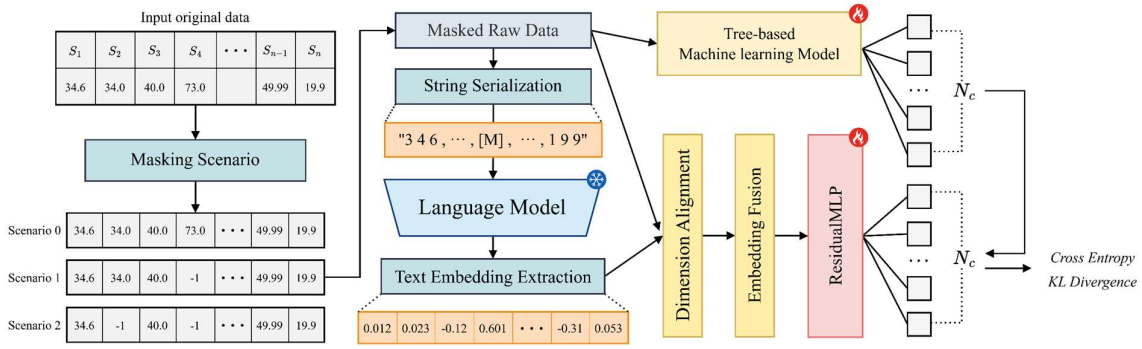

Fig. 1. Overview of a Hybrid architecture with a machine learning teacher and a Language Model Embedding-based student (HyLME).

outliers, the inter-variable relationships can be preserved, and similar classes are expected to be projected into a shared embedding space. However, when applied to domains not covered by the pre-trained model's knowledge base, a lack of domain-specific knowledge may occur, and the degradation in representational capacity can become more pronounced.

To address this limitation, we incorporate XGBoost as a complementary model. XGBoost has a structure that learns by repeatedly partitioning the data into intervals, and it possesses a robust characteristic whereby the overall model structure and predictive performance are not significantly affected even in the presence of outliers or missing data in the input data.

Accordingly, as shown in Fig. 1, this study adopts a knowledge distillation architecture in which the embeddings from the language model are fused with the original data as input, and the prediction signals from the tree-based machine learning model are used as supervised learning targets to learn the decision process. This approach is designed to precisely capture the relationships among variables in complex sensor data and to perform predictive maintenance robustly in the presence of missing data.

B. *Extracting Text Embeddings from Pre-trained Language Models*

Masking Scenario: In this study, a masking scenario is introduced in which a predetermined number of variables are forcibly set to "-1" according to each scenario, indicating missing values or sensor outliers and thereby simulating a missing data environment. This simulates real-world sensor failure patterns that maintenance systems frequently encounter. This method is designed to train the model's generalization performance under missing data conditions and to verify its robustness.

Similarly, Caruso et al. [12] applied epoch-wise random masking to missing data, but the per-step resampling introduced additional computational cost. To mitigate this overhead, we randomly generate a single masking pattern at initialization and reuse it for all training iterations.

This masking scenario operates differently between the training and testing phases. In the training phase, masking is applied randomly, whereas during testing variables are masked sequentially in descending order of XGBoost-derived feature importance so that we can directly observe the impact of losing the most critical sensors. The number of masked sensors in each scenario increases consistently as the scenario level increases.

String Serialization: The sensor data preprocessed according to the masking scenario is used to extract text embeddings through a pre-trained language model (BERT). As shown in Fig. 1, the data input to the language model is converted into a string based on the digit-space tokenization strategy.

This method, devised in the study by Kaur et al. [14], converts numerical values into integers at a precision of one decimal place and then represents each digit by separating them with spaces. For example, the value "49.99" is replaced with "4 9 9".

In addition, the values between sensors are separated by commas, and the masked value "-1" is replaced with a special token "[M]". The final constructed string is then converted into a token sequence through the pre-trained tokenizer of the language model and used for embedding extraction.

Text Embedding Extraction: The text serialized through the digit-space tokenization and special tokens is used to extract text embeddings via a language model. In this study, embeddings are extracted using a pre-trained BERT, and the embedding extraction process is inspired by Tang et al. [13].

Given an input string x, it is converted into a token sequence of length L, denoted as $T = T(x) = (t_1, t_2, \dots, t_L)$, using a pre-trained tokenizer. This token sequence is propagated through the Transformer structure of the language model, resulting in a hidden state $H = [h_1, h_2, \dots, h_L]^T \in \mathbb{R}^{L \times d}$, where d is the embedding dimension of the BERT (768). The final text embedding is obtained through mean pooling, as expressed in (1).

$$\phi(T) = \frac{1}{L}\sum_{i=1}^{L} h_i \in \mathbb{R}^d \tag{1}$$

C. *Machine Learning Teacher and Deep Learning Student*

The extracted text embedding is fused with the original data, where the smaller-dimensional input is zero-padded to match the larger dimension for alignment. Here, the fusion is performed through element-wise addition. This method is inspired by the study of Kaur et al., which reported that simple addition was more effective than concatenation. Following this approach, the present study also constructs the input data by adopting the fusion method of addition after dimensional alignment.

This data is used as input to the ResidualMLP head for classification, and the training is performed using a knowledge distillation approach that learns the logits of XGBoost. In this architecture, the ResidualMLP acts as the student and XGBoost as the teacher, with the student learning from the knowledge of the teacher. XGBoost is used after being pre-

trained on the corresponding dataset using the original input rather than the embedding. The loss function for training the ResidualMLP can be briefly represented as follows. Based on empirical evaluation, the hyperparameter T in (3) was selected to be 4.0, and λ in (4) was selected to be 0.5.

$$L_{CE} = Cross\ Entropy(y, \hat{y}) \tag{2}$$

$$L_{KD} = T_s^2 \cdot KL(softmax\left(\frac{z_t}{T_t}\right) || \ softmax\left(\frac{z_s}{T_s}\right)) \tag{3}$$

$$L_{total} = (1-\lambda) \cdot L_{CE} + \lambda \cdot L_{KD} \tag{4}$$

IV. Experiments

A. Datasets

To validate the predictive maintenance accuracy and robustness to missing data of the proposed architecture, the Multimodal Data for Predictive Maintenance of Transport Devices in Manufacturing Sites (MPTMS) [17], which consists of 13,121 sample sets collected from real-world manufacturing environments, was used. The MPTMS dataset consists of sensor data and thermal images for carbonization prediction, applicable in real manufacturing sites for semiconductors, displays, and automobiles. It is provided with labels for a total of four conditions, which are normal, caution, warning, and danger conditions.

From this dataset, excluding the image data, data from eight types of sensors were extracted and used. The sensor data was restructured into a tabular format for processing. The sensors used include NTC, PM1.0, PM2.5, PM10, CT1, CT2, CT3, and CT4, and their correlations with the classification labels are shown in Fig. 2. In particular, a high degree of correlation was observed among the PM-series sensors, and a certain level of interdependence was also found among the CT-series sensors.

The feature importance of the entire sensor dataset was calculated using the XGBoost model, and the results are presented in Fig. 3. A notable finding is that there is a significant drop after the top three sensors, suggesting that a small subset of sensors has a dominant impact on prediction performance. In this study, masking scenarios were applied to the MPTMS dataset, and comparative experiments were conducted by training both baseline models and HyLME.

B. Baseline Competitive Models

The baseline models were constructed using both methodologies adopted in previous studies and widely recognized SOTA models, Linear Regression [18], XGBoost [16], and Support Vector Machine (SVM) [19] are machine learning-based models that have demonstrated strong performance on tabular data. MLP [20] and ResidualMLP, which incorporates the idea of residual connections from ResNet [21], are basic deep learning model whose performance has already been proven in many fields. In addition, AutoEncoder [10], an unsupervised model that learns to compress input data into a latent space and reconstruct it, has been widely used as an approach for generating missing data. Finally, Transformer-based models known to achieve SOTA performance on tabular data, such as TabNet [22], TabTransformer [23], and FTTransformer [24], were also included for comparative experiments.

C. Evaluation Metrics

The evaluation metrics for the experimental results are accuracy, precision, recall, and F1-Score, and the formulas for each are defined in (5)-(8).

$$Accuracy = \frac{\sum_{i=1}^{C} TP_i}{\sum_{i=1}^{C}(TP_i + FP_i + FN_i)} \tag{5}$$

$$Precision_i = \frac{TP_i}{TP_i + FP_i}, \quad Precision_{macro} = \frac{1}{C}\sum_{i=1}^{C} Precision_i \tag{6}$$

$$Recall_i = \frac{TP_i}{TP_i + FN_i}, \quad Recall_{macro} = \frac{1}{C}\sum_{i=1}^{C} Recall_i \tag{7}$$

$$F1_{macro} = \frac{1}{C}\sum_{i=1}^{C} \frac{2 \times Precision_i \times Recall_i}{Precision_i + Recall_i} \tag{8}$$

Fig. 2. Correlation heatmap of the Multimodal Data for Predictive Maintenance of Transport Devices in Manufacturing Sites (MPTMS).

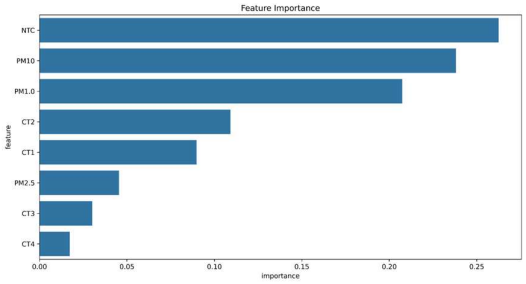

Fig. 3. Feature importance of sensors in the MPTMS datasets.

D. Results

Quantitative Evaluation: As shown in Table 1, the proposed HyLME achieved the highest performance among all comparison models, with an average scenario accuracy of 0.83207. Although the improvement over XGBoost, the second-best model, was relatively modest at approximately 0.36%, this difference becomes more substantial in scenarios with multiple missing sensors. Specifically, the performance gains became more noticeable from Scenario 4 onward, where missing data began to affect the top three sensors identified by feature importance.

The key point here is that, in Scenarios 4 through 7, HyLME achieved higher accuracies than XGBoost by 1.241%, 0.383%, 0.481%, and 0.691%, respectively, with an average improvement of 0.699%, exceeding the overall average gain. These results suggest that HyLME presents a meaningful advancement in addressing the limitations of existing models under missing data conditions.

Despite being a model renowned for achieving SOTA performance in numerous studies thanks to its robustness against outliers and missing data, XGBoost exhibited a decli ne in performance as the number of missing sensors grew. In

TABLE I. QUANTITATIVE STUDY RESULTS

Models	Accuracy of Masking Scenario								
	0	1	2	3	4	5	6	7	Average
LinearRegression	0.79934	0.49411	0.49580	0.49508	0.46345	0.45530	0.45530	0.45530	0.51421
XGBoost	0.93392	**0.92827**	**0.92642**	0.92585	0.81983	0.79805	0.73850	0.56180	0.82908
SVM	0.87470	0.85501	0.84589	0.84678	0.64491	0.61433	0.56261	0.50420	0.71855
MLP	0.90036	0.88866	0.88091	0.87010	0.71752	0.71284	0.62667	0.50347	0.76257
ResidualMLP	0.91932	0.91189	0.90116	0.88212	0.79377	0.75472	0.69590	0.54543	0.80054
AutoEncoder (R)	0.85162	0.68372	0.53195	0.52945	0.31023	0.30353	0.31370	0.16242	0.59213
AutoEncoder (L)	0.84622	0.84008	0.81548	0.78425	0.67872	0.63555	0.61578	0.49177	0.71348
TabNet	0.88761	0.87801	0.87768	0.87502	0.72906	0.70179	0.54591	0.47684	0.74649
TabTransformer	0.85646	0.86566	0.82072	0.77376	0.59981	0.63684	0.50258	0.45530	0.68889
FTTransformer	0.91730	0.91496	0.90963	0.89285	0.81459	0.77441	0.69905	0.53929	0.80776
HyLME (Ours)	**0.93424**	0.92730	0.92609	**0.93005**	**0.83000**	**0.80111**	**0.74205**	**0.56568**	**0.83207**

TABLE II. ABLATION STUDY RESULTS

Models	Average of Masking Scenario			
	Acc	P	R	F1
MLP	0.76257	0.79902	0.70758	0.71509
ResidualMLP	0.80054	0.81316	0.75925	0.76028
ResidualMLP (KD)	0.81995	0.84366	0.77813	0.77989
HyLME (w/o Fusion)	0.81867	0.81532	0.77281	0.77361
HyLME (w/o KD)	0.83189	0.84534	**0.79445**	**0.80002**
HyLME (Ours)	**0.83207**	**0.85423**	0.78957	0.79723

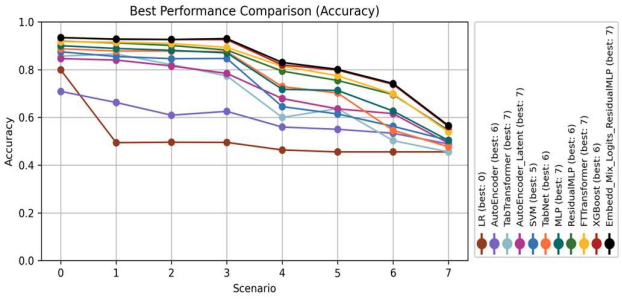

Fig. 4. Scenario-wise Accuracy Comparison of Baseline Models under Missing Sensor Conditions.

contrast, our proposed model not only attained the highest initial performance but also delivered the top accuracy in all later scenarios, with the exception of Scenarios 1 and 2.

Furthermore, the proposed model showed 3.01% higher accuracy than FTTransformer, a known SOTA model for tabular data that has a transformer structure similar to language models. In conclusion, our model recorded an average accuracy 15.99% higher than the average accuracy of 0.71737 from the other models in the comparison group, excluding HyLME. This performance improvement supports the proposed architecture's capability to handle missing data environments and demonstrates its effectiveness in enhancing both accuracy and robustness. Here, Fig. 4 represents the scenario-wise accuracy of each model, and the results presented in Table 1 and Fig. 2 were recorded by selecting the model with the highest average accuracy in each scenario.-

Ablation Study: The HyLME architecture assumes that the pretrained language model, with its context-aware learning capabilities, can capture relationships among sensors and project instances belonging to the same class closer together in the embedding space. However, since it was not trained on the downstream task, the model aimed to alleviate the issue of knowledge deficiency by learning the logits from the tree-based XGBoost model, which is robust to missing data.

To demonstrate the effect of each component on model performance, an ablation study was conducted, with the results shown in Table 2. Primarily, the ResidualMLP was chosen over a MLP to mitigate the vanishing gradient problem and ensure stable training via residual connections. Based on accuracy, the ResidualMLP recorded a score of 0.80054, representing a 4.98% improvement compared to the MLP.

When the text embedding from the language model was applied, an accuracy of 0.81867 was achieved, which is 2.26% higher than the result without it. Additionally, when the text embedding was fused with the original data, the performance improved by 3.92% compared to the standard ResidualMLP, demonstrates that the fusion led to better results. This indicates that training on the fused embeddings from the pre-trained language model is effective for improving performance.

In addition, an accuracy of 0.81995 was achieved by training with knowledge distillation from XGBoost without incorporating language model embeddings, confirming the effectiveness of learning from a tree-based model. However, when the language model embeddings were first fused and then used in the knowledge distillation process, the accuracy improved to 0.83207, representing a 1.48% enhancement. These experimental results demonstrate that each component proposed in the HyLME architecture contributes effectively to improving predictive maintenance performance under missing data conditions.

V. CONCLUSIONS

In this paper, we proposed HyLME, a novel approach for robust predictive maintenance under missing sensor conditions. HyLME integrates text embeddings from pre-trained language models with tree-based machine learning through knowledge distillation, effectively combining the contextual understanding capabilities of BERT with the robustness of XGBoost. Our approach serializes sensor data

into text format using digit-space tokenization and extracts embeddings that capture inter-sensor relationships, which are then fused with original features and processed through a ResidualMLP student network guided by XGBoost teacher logits.

Comprehensive experiments on the MPTMS dataset with 13,121 real-world manufacturing samples demonstrated HyLME's effectiveness. The proposed method achieved an average accuracy of 83.21% across all missing data scenarios, outperforming ten baseline models by 15.99% on average. Notably, HyLME maintained superior performance stability as sensor failures increased, showing particular strength in scenarios where critical sensors were missing (maintaining over 80% accuracy with up to five missing sensors). The ablation study revealed that language model embeddings alone contributed a 14.74% improvement over traditional autoencoder approaches, validating our hypothesis that pre-trained language models can effectively capture complex sensor relationships even in degraded conditions.

Despite these promising results, we acknowledge several limitations. First, the knowledge distillation approach inherently bounds student performance by teacher capabilities, which may explain why HyLME's performance closely tracks XGBoost in certain scenarios. Second, our current approach does not explicitly model temporal dependencies in sensor data, focusing instead on inter-sensor relationships at individual time points. Third, computational overhead from text serialization and embedding extraction may limit real-time deployment in resource-constrained environments.

Future research will address these limitations through several directions: (1) exploring alternative knowledge transfer mechanisms that allow the student to surpass teacher performance, (2) incorporating temporal modeling through recurrent or attention-based architectures while maintaining robustness to missing data, and (3) validating generalization across diverse industrial datasets and sensor modalities. Additionally, we plan to analyze the learned embeddings to better understand which linguistic patterns correspond to sensor relationships, potentially leading to more interpretable predictive maintenance systems.

VI. Acknowledgement

This work was supported by the Industrial Technology Innovation Program (RS-2025-02633048, To improve the competitiveness of the manufacturing industry, the language model for defense aviation and the development of Korean-style design automation model technology with twice the speed of standard parts design) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

References

[1] M. Samanazari, F. Flammini, S. Santini, and M. Caporuscio, "A systematic literature review on transfer learning for predictive maintenance in Industry 4.0," IEEE Access, vol. 11, pp. 12887–12910, Jan. 2023, doi:10.1109/ACCESS.2023.3239784.

[2] T. Zonta et al., "Predictive maintenance in the Industry 4.0: A systematic literature review," Comput. Ind. Eng., vol. 150, Art. no. 106889, 2020, doi:10.1016/j.cie.2020.106889.

[3] C. M. Carbery, R. Woods, C. McAteer, and D. M. Ferguson, "Missingness analysis of manufacturing systems: a case study," Proc. Inst. Mech. Eng. B, vol. 236, no. 10, pp. 1406–1417, 2022.

[4] P. Nunes, J. Santos, and E. Rocha, "Challenges in predictive maintenance–A review," CIRP J. Manuf. Sci. Technol., vol. 40, pp. 53–67, 2023.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. 2019 Conf. North Am. Chapter Assoc. Comput. Linguistics Hum. Lang. Technol., vol. 1, Jun. 2019, pp. 4171–4186, doi:10.18653/v1/N19-1423.

[6] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," OpenAI Blog, vol. 1, no. 8, p. 9, 2019.

[7] A. Grattafiori et al., "The llama 3 herd of models," arXiv preprint arXiv:2407.21783, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.21783.

[8] A. P. Hermawan, D. S. Kim, and J. M. Lee, "Predictive maintenance of aircraft engine using deep learning technique," in Proc. 2020 Int. Conf. Information and Communication Technology Convergence (ICTC), Oct. 2020, pp. 1296–1298.

[9] A. P. Ghadekar, A. Manakshe, S. Madhikar, S. Patil, M. Mukadam, and T. Gambhir, "Predictive maintenance for industrial equipment: Using XGBoost and local outlier factor with explainable AI for analysis," in Proc. 2024 14th Int. Conf. Cloud Comput., Data Sci. & Eng. (Confluence), Jan. 2024, pp. 25–30.

[10] A. H. Ba-Alawi, J. Loy-Benitez, S. Kim, and C. Yoo, "Missing data imputation and sensor self-validation towards a sustainable operation of wastewater treatment plants via deep variational residual autoencoders," Chemosphere, vol. 288, Art. no. 132647, 2022.

[11] J. Lee, Y. Yu, and H. Seo, "Restoration of multi-channel signal loss using autoencoder with recursive input strategy," Sci. Rep., vol. 15, no. 1, Art. no. 13729, 2025.

[12] C. M. Caruso, P. Soda, and V. Guarrasi, "Not another imputation method: A transformer-based model for missing values in tabular datasets," unpublished; submitted to AI Open, 2024. [Online]. Available: https://arxiv.org/abs/2407.11540.

[13] E. Tang, B. Yang, and X. Song, "Understanding LLM embeddings for regression," Trans. Mach. Learn. Res., Feb. 2025. [Online]. Available: https://openreview.net/forum?id=Wt6Iz5XNIO.

[14] R. Kaur, Z. Zeng, T. Balch, and M. Veloso, "LETS-C: Leveraging Text Embedding for Time Series Classification," in Proc. NeurIPS Workshop on Time Series in the Age of Large Models, 2024.

[15] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, submitted to NIPS 2014 Deep Learning Workshop, Mar. 2015. [Online]. Available: https://arxiv.org/abs/1503.02531.

[16] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, Aug. 2016, pp. 785–794.

[17] HCI Plus Co., "Multimodal Data for Predictive Maintenance of Transport Devices in Manufacturing Sites," AI Hub, National Information Society Agency (NIA), dataset no. 71802, Dec. 20 2024. [Online]. Available: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=71802. Accessed: Jun. 14 2025.

[18] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.

[19] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Schölkopf, "Support vector machines," IEEE Intell. Syst. Appl., vol. 13, no. 4, pp. 18–28, Jul.–Aug. 1998, doi:10.1109/5254.708428.

[20] H. Ramchoun, Y. Ghanou, M. Ettaouil, and M. A. Janati Idrissi, "Multilayer perceptron: Architecture optimization and training," Int. J. Interact. Multimedia Artif. Intell., vol. 4, no. 1, Sept. 2016, doi:10.9781/ijimai.2016.415.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 770–778.

[22] S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," in Proc. AAAI Conf. Artif. Intell., vol. 35, no. 8, May 2021, pp. 6679–6687.

[23] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "TabTransformer: Tabular data modeling using contextual embeddings," arXiv preprint arXiv:2012.06678, Dec. 2020. [Online]. Available: https://arxiv.org/abs/2012.06678.

[24] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 18932–18943.