

## RESEARCH ARTICLE

# Transformer-Based Fault Detection Using Pressure Signals for Hydraulic Pumps

**A. RAN KIM<sup>1</sup>, HA SEON KIM<sup>1</sup>, AND SUN YOUNG KIM<sup>2</sup>, (Member, IEEE)**<sup>1</sup>Department of Mechanical Engineering, Kunsan National University, Gunsan 54150, Republic of Korea<sup>2</sup>School of Mechanical Engineering, Kunsan National University, Gunsan 54150, Republic of Korea

Corresponding author: Sun Young Kim (sykim77@kunsan.ac.kr)

This work was supported by Korea Foundation for Women in Science, Engineering and Technology (WISSET) grant funded by the Ministry of Science and ICT (MSIT) under the team research program for female engineering students and supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0012769, The Competency Development Program for Industry Specialist). Also, this research was supported by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT, the South Korea (No. 2021R1C1C1009219).

**ABSTRACT** In this paper, modified transformer-based fault detection for a hydraulic pump is performed using the pressure signals of the hydraulic pump. The pump is considered a swash plate axial piston pump used in the excavator. Additionally, the outlet pressure data of the pump are extracted based on Amesim. The proposed transformer is a modified transformer, which allows fast fault detection by modifying the transformer and reducing the size of this model. The classes are normal and 6 fault types, and comparison models are long short-term memory (LSTM) and its family models, which are representative time series models. Unlike comparison models, the modified transformer has an average accuracy of 100% and a detection time of 0.00271 s, which is a slight difference of 0.00036 s from the single LSTM that showed the shortest operation time among the models. We also perform fault detection by changing data points and show a stable high accuracy of 99.93% for all data points of 500, 1,000, and 1,500 without any optimization. Various external noises are added because excavators are construction equipment used in rough terrain. Therefore, we conduct detection performance analysis at different additional noise levels with Gaussian noise with zero mean. As a result, we confirm that the modified transformer showed a high detection accuracy of over 98.08% up to standard deviation 4, where data characteristics were well maintained, unlike other time series models. Through the various analyses above, we confirm that fast and accurate fault detection is possible based on the modified transformer.


**INDEX TERMS** Hydraulic pump, pressure sensor, fault detection, transformer, long short-term memory, attention mechanism.

## I. INTRODUCTION

An excavator is a construction machine that requires tremendous power and is operated by a hydraulic system. A hydraulic system uses hydraulic energy generated by compressing hydraulic fluid to control moving parts such as hydraulic cylinders and motors. It provides constant force and torque, allowing large power transfers and easy control of force, direction, and speed for safe and precise control. Hydraulic systems are difficult to identify and categorize faults, and more than 50% of faults are

associated with pumps [1]. Pumps can fail because of various problems, including slugging, excessive heat, worn swash plates, and clogged hoses. If these problems are not resolved immediately, these faults can cause substantial production interruptions, resulting in economic losses and jeopardizing the safety of employees operating the equipment. Therefore, accurate and rapid detection of hydraulic pump faults is important to ensure the reliability and economic feasibility of equipment and hydraulic systems.

Hydraulic pump faults can be detected using two methods. System-based fault detection, which has been primarily used in the past, is based on mathematical models and can detect faults with high accuracy and a small amount of

The associate editor coordinating the review of this manuscript and approving it for publication was Mehrdad Saif .

data. However, the system-based fault detection of hydraulic pumps is difficult to use when high accuracy is required due to the hard implementation. Although the variables of the system have a remarkable impact on the detection accuracy, they must be selected as per user requirements. Additionally, the selection of thresholds determines the performance of the fault detection system. However, it is difficult to select appropriate thresholds without expertise [2], [3]. To address these problems, artificial intelligence (AI)-based fault detection has been actively studied, which does not require the setting of variables or thresholds, unlike existing system-based methods. For this reason, AI-based fault detection has been widely used recently.

For hydraulic pumps, fault detection was mostly based on vibration signals, whereas other sensor data have hardly been used [4]. Among hydraulic pumps, axial piston pumps, which are commonly used in excavators, also use vibration signals in most cases. Moreover, there are two main types of studies: those based on support vector machine (SVM), i.e., machine learning, and those based on two-dimensional convolutional neural network (2D CNN), i.e., deep learning. In the following, we first present studies on fault detection of piston pumps based on vibration signals using SVM. Casoli et al. [5] confirmed the feasibility of fault detection in axial piston pumps based on machine learning such as SVM and the k-nearest neighbor algorithm. Tian et al. [6] combined the signal processing algorithms wavelet packet transform, singular value decomposition, and SVM to perform fault diagnosis on a small number of datasets. Tang et al. [7] performed 2D CNN-based fault detection of swash plate axial piston pumps and found that the average accuracy of the proposed CNN was as high as 98.44%. To generate input data for the 2D CNN, the original vibration signal was converted into an image using a continuous wavelet transform (CWT), considering normal and four faults. Furthermore, Zhu et al. [8] proposed a new intelligent fault detection method, namely, CWT-AlexNet, which combines CWT and 2D CNN and considers normal and four faults. A previous study [9] applied batch normalization layer and particle swarm optimization to LeNet for intelligent fault detection of hydraulic piston pumps and reported a high accuracy of 98.71% and lower test time compared to other models. Furthermore, 2D CNN has been used in deep learning-based fault detection for hydraulic pumps [10], [11], [12].

Studies have been conducted on the detection of piston pump faults based on pressure signals using AI. Tang et al. [13] proposed an adaptive deep CNN that automatically optimizes hyperparameters, which are the configuration variables of the model, by applying a model that combines Bayesian optimization and 2D CNN. Furthermore, the multisignal-based fault detection of piston pumps has been investigated using 2D CNN. Tang et al. [14] proposed a model combining batch normalization and Bayesian optimization for 2D CNN. They considered normal and four faults and used vibration, pressure, and acoustic signals as

inputs. By applying the proposed model, the detection accuracy increased by 11.06% compared to the existing LeNet5. In addition to piston pumps, fault detection [15], [16] was performed for other hydraulic pumps such as vane and centrifugal pumps.

However, the most commonly used 2D CNN is computationally intensive because it uses 2D images as input. Besides, using one-dimensional (1D) data, such as a signal over time, requires an image conversion process. For the aforementioned reasons, 2D CNN is unsuitable for rapid fault detection. In the field of fault detection other than hydraulic pumps, time series models with 1D signals as input have been used to detect motor faults [17] and bearing faults [18], aircraft hydraulic system faults [19]. Furthermore, efforts have been made to apply a transformer, a model that performs exceptionally in machine translation tasks, to the fault detection field instead of CNN and recurrent neural network (RNN), which are popular models for fault detection [20], [21], [22], [23]. However, AI-based detection of piston pump faults is less explored than fault detection for bearings, motors, wear, and rotating machinery. In particular, only a few studies have focused on time series model-based fault detection, and there is no research on transformer-based fault detection.

Recently, research [24], [25], [26] has been conducted to reduce the model size by applying methods such as signal quantization and structure changes of the transformer. However, no research has been conducted on reducing the size of the transformer in pump fault detection.

Therefore, in this paper, we perform deep learning-based fault detection using the pressure signal of a swash plate axial piston pump extracted based on simulation and propose an appropriate deep learning model through performance analysis. First, we construct pressure data of the pump that reflect reality through the data preprocessing because simulation-based data do not consider the operating principle of the pump and external noise such as the vibration of the machine itself. Deep learning models consider LSTM and BiLSTM, which are still widely used for prediction and classification as representative time series models. Here, BiLSTM is a model that applies the bidirectional concept of LSTM. Furthermore, we applied attention mechanism (AM), the main idea of the transformer, to LSTM and BiLSTM (LSTM AM and BiLSTM AM, respectively), and explored these models for fault detection. Additionally, we selected the transformer, which can be considered the fundamental of ChatGPT, as the detection model and proposed a modified model for pump fault detection. Moreover, we reduced the size of the transformer and LSTM family models since using heavy models is expensive in the actual industrial field. The main contributions can be summarized in threefold.

- 1) Using pressure signals rather than vibration signals which is widely used for deep learning-based hydraulic pump fault detection.

- 2) For real-time fault detection using one-dimensional pressure signals, selecting time series models rather than the 2D CNN that was widely used previously and optimizing models by considering fewer parameters than existing papers to make models lighter.
- 3) Designing a model with high accuracy and short fault detection time by modifying the transformer, which is widely used in translation problems among time series models and has not been previously used for hydraulic pump fault detection.

The remainder of this paper is organized as follows. Section II describes the introduction to the dataset and building the dataset introduces the theoretical background of the vanilla transformer and its modified structure for fault detection in hydraulic pumps. Section III. Section IV discusses the optimization process of selected 5 models and presents the optimization results. Section V analyzes the performance of the optimized models and conducts performance comparisons and analysis depending on the noise and the number of data points. Finally, we summarize the analysis results in Section VI and conclude with suggestions for future studies on improving the fault detection performance.

## II. DATASET

In this study, the hydraulic pump considered is a variable displacement swash plate axial piston pump used in 20-ton excavators. Also, realistic training and validation datasets are critical for AI-based fault detection. Excavators perform tasks such as digging the ground and breaking stones at construction sites, so various noises are added to the output signal of the pump used in the excavator. However, pressure data extracted through simulation do not reflect the operating principle of the pump and external noise. Thus, the process of constructing a realistic dataset is thoroughly described in this section by simulating real-world situations.

### A. SELECTION OF PUMP AND FAULT TYPE

The pump that we considered is a swash plate axial piston pump. There are various types of hydraulic pumps with different characteristics and operating principles. The pump considered here is a swash plate axial piston pump, which belongs to the category of piston pumps among the variable displacement hydraulic pumps, whose drive shaft varies the discharge volume per revolution. A piston pump operates a hydraulic system by reciprocating the piston motion in a cylinder block and generating hydraulic pressure. The selected pump is characterized by the following features: high-speed operation, high power, and complex structure. The specific working principle is as follows: as the shaft rotates, the piston and cylinder rotate together while the piston performs a reciprocating motion in the cylinder. The hydraulic pressure is generated by introducing and discharging the working fluid into the port on the valve plate through the reciprocating motion [27].

In previous research [7], [8], [14], fault detection mainly considered four types of faults. However, fault detection becomes more difficult as the number of types of faults increases, and different types of faults occur in a real environment. Therefore, we performed fault detection of a hydraulic pump considering Normal and six faults. Specifically, to the three faults investigated in the previous study [28] (piston port blockage, hose blockage, and hose explosion), we added three faults caused by friction (friction between cylinder and valve plate, piston and cylinder, and shoe and swash plate). The Abbreviations of each class are summarized in Table 1.

**TABLE 1. The information of classes for hydraulic axial piston pump under 7 conditions.**

Abbreviation	Description
Normal	Normal
HosBlk	Blockage of hose
HosExp	Explosion of hose
PstBlk	Blockage of one piston
PstFrc	Friction between piston and cylinder barrel
SSWFrc	Friction between shoe and swash plate
BVPFrc	Friction between cylinder barrel and valve plate

### B. GENERATION OF PUMP DATA USING SIMULATOR

Data are based on Amesim, which extracts outlet pressure data by class at a frequency of 5,000 Hz. Orifice diameter, outlet volume, and rotation speed are set to 13.00 mm, 20.00 cm<sup>3</sup>, and 2,600 rpm, respectively, to extract pressure data over time. The data extracted here are data with a fixed angle of the swash plate. Thus, we extracted 8 pressure data over time per class with different angles of the swash plate considering the appropriate operating angles from 2.4° to 16.4° in 2° increments.

### C. DATA PREPROCESSING

#### 1) CONVERTING TO DYNAMIC PUMP DATA

However, the selected pump has a variable swash plate angle, whereas the data extracted from the simulation are pressure data with a fixed swash plate angle. Accordingly, we used the 8 datasets extracted by class to extract the pressure data as the swash plate angle changes over time so that the data reflect the operating principle of the actual pump. The result of the extraction is one dynamic pressure data per class.

First, patterns were visualized to analyze the extracted dynamic data as shown in Fig. 1. In the figure, the data for normal (Fig. 1(a)) and friction-induced faults (Fig. 1(e)-(g)) have remarkably similar geometries and pressure magnitudes. Among these, the geometries and pressure magnitudes of Normal(Fig. 1(a))-PstBlk(Fig. 1(e)) and SSWFrc(Fig. 1(f))-BVPFrc(Fig. 1(g)) are difficult to classify through the naked eye. Therefore, we estimate that classification for Normal and friction-induced faults is difficult, and especially Normal-PstBlk and SSWFrc-BVPFrc will be unclassified among friction-induced faults.

Additionally, to quantify the visual similarity of patterns, similarity tests are performed patterns as shown in Table 2. The similarity test uses the structural similarity index measure (SSIM) [29] used in image evaluation techniques. This technique is an evaluation metric by utilizes that human vision is sensitive to structural information in images and has been evaluated as more similar to the human visual system than the peak signal-to-noise ratio (PSNR) which is a commonly used image evaluation metric. Also, SSIM is a quantified value that compares the luminance, contrast, and structure of two images. The values extracted through SSIM range from 0 to 1 and are closer to 1 indicating that the two images are visually similar. In this study, to reduce the effects of luminance and contrast, we conduct a similarity test after converting the images of Fig. 1 to grayscale. In the table, similar to the visual analysis results above, there is a high level of similarity between the signal patterns for normal and friction-induced faults. Furthermore, we judged that classification is difficult because the similarity of SSWFrc-BVPFrc and Normal-PstFrc, which showed the highest similarity, is 0.9932 and 0.9912, respectively.

## 2) DATA AUGMENTATION

The dynamic pressure data of the hydraulic pump extracted from Section II-C1 reflects reality, but there is only one per class. However, a large amount and high quality of data are required for applying AI. Hence, we perform data augmentation to construct datasets for AI training. Data augmentation is a technique used in the field of computer vision to solve the problem of poorly trained models due to insufficient training data. Later, data augmentation techniques are also applied to time series data to solve the problem of sparse data.

A time series data are collection of observations obtained through repeated measurements over time and has essential characteristics such as irregularity, cyclicity, trend, and seasonality. To maintain and utilize the inherent characteristics of such time series data, techniques such as rotation commonly used in existing images cannot be used [30]. Therefore, to preserve the characteristics of time series data, data augmentation was performed using the existing algorithm and generative adversarial network (GAN) in previous

studies [31], [32], [33]. For GAN-based data augmentation, data for GAN training must exist. However, GAN-based data augmentation techniques are not suitable for use because there is only one data per class. Another data augmentation technique, traditional methods, are easy to implement and mainly used for time series as shown in Fig. 2.

In this study, data augmentation is performed by applying jittering among traditional methods. Jittering is a technique frequently used on time series data by adding noise to the data. In particular, jittering is often used for sensor-based extracted time series data. Our simulation-based sensor data are noise-free, but a lot of external noise comes into the actual measurement sensor data of excavators used in rough terrain. Thus, we consider jittering among traditional data augmentation techniques to construct a dataset that reflects reality. At this time, the noise is assumed to be Gaussian noise with zero mean. Additionally, vibration of various levels occurs in real-world situations. Additionally, a previous study [34] showed higher performance of noise-robust at AI-based fault detection by constructing the training dataset with various noise levels than when considering a single noise. Therefore, to simulate real-world situations and construct the noise-robust training dataset, we consider standard deviations to vary from 1 to 10. We extract 1,000 dynamic pump pressure datasets per class through a selected data augmentation technique.

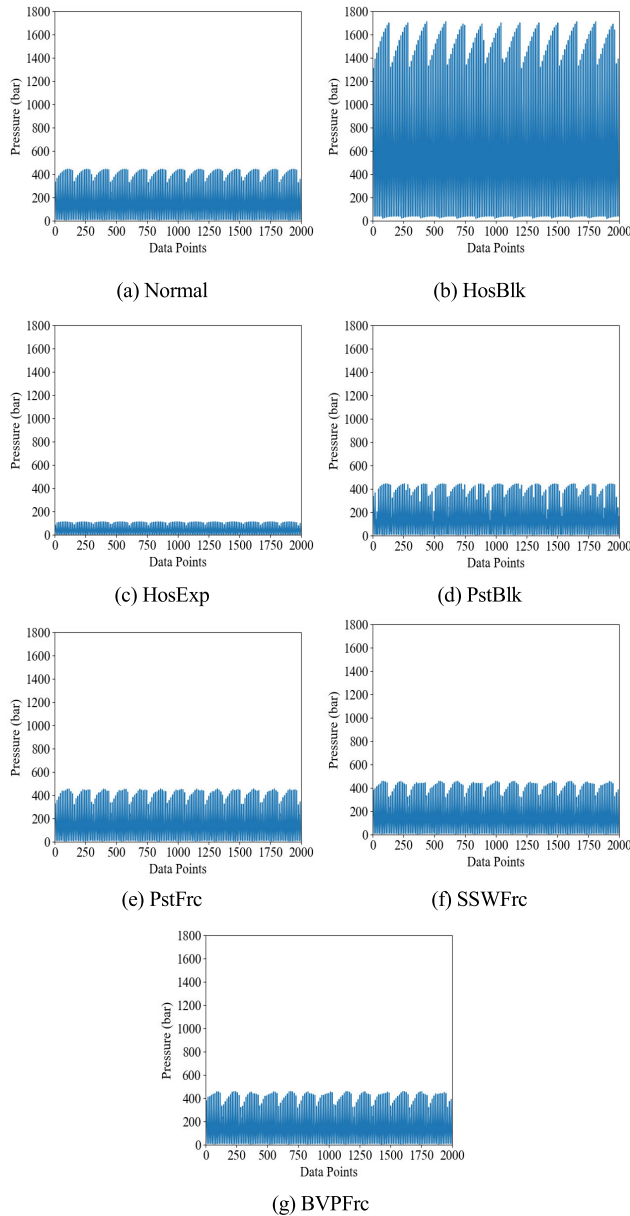
## D. CONSTRUCTION OF TRAINING AND VALIDATION DATASET

To properly verify the AI technique, the data splitting ratio is also important. At this time, if the number of constructed datasets is large such as hundreds of thousands, the ratio of training and test datasets does not matter. So, the ratio of 99:1 has no problem with proper verification. However, test datasets of high quality are required for accurate validation when the number of datasets is small. Moreover, a K-fold cross-validation method also exists, but it is difficult to apply in this study because network optimization is performed by considering various hyperparameters and comparing performance. Thus, since this paper belongs to small-scale datasets, we perform data splitting at a fixed ratio of 7:3. That is, of the

TABLE 2. Results of similarity test.

	(a) Normal	(b) HosBlk	(c) HosExp	(d) PstBlk	(e) PstFrc	(f) SSWFrc	(g) BVPFrc
(a) Normal	1.0000	0.3995	0.8357	0.9895	<b>0.9912</b>	0.9742	0.9742
(b) HosBlk	0.3995	1.0000	0.4244	0.4016	0.3985	0.3992	0.3993
(c) HosExp	0.8357	0.4244	1.0000	0.8373	0.8350	0.8340	0.8339
(d) PstBlk	0.9895	0.4016	0.8373	1.0000	0.9822	0.9656	0.9658
(e) PstFrc	<b>0.9912</b>	0.3985	0.8350	0.9822	1.0000	0.9742	0.9743
(f) SSWFrc	0.9742	0.3992	0.8340	0.9656	0.9742	1.0000	<b>0.9932</b>
(g) BVPFrc	0.9742	0.3993	0.8339	0.9658	0.9743	<b>0.9932</b>	1.0000





**FIGURE 1.** Dynamic pressure data of hydraulic pump according to classes.

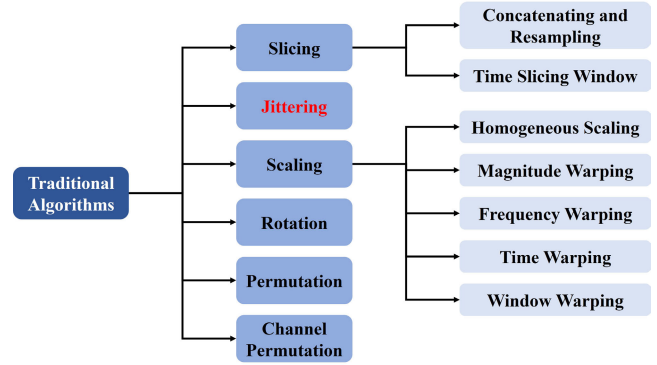
1,000 dynamic pump pressure data extracted for each class, 700 are used for training and 300 are used for validation.

### III. TRANSFORMER

This section describes the vanilla transformer and introduces the modified transformer.

#### A. VANILLA TRANSFORMER

RNN [35] and LSTM [36], representative time series data processing models, receive and process input sequences serially and cannot be computed in parallel, which reduces the computing speed. Hence, an AM [37]-based model, the vanilla transformer, was proposed in 2017 [38]. The vanilla transformer eliminates the iterative structure of traditional



**FIGURE 2.** Classification of traditional time series data augmentation methods.

time series models and uses a multi-head AM to perform parallel computations.

The vanilla transformer comprises an encoder and a decoder, similar to sequence-to-sequence (seq2seq) [39]. It has a multilayer structure with an encoder block that extracts the features of the input sequence and a decoder block that generates the output sentence based on the information extracted by the encoder block (Fig. 3) [40]. A description of the core layers considered in this paper is as follows.

#### 1) POSITIONAL ENCODING [38], [41]

Unlike RNN and LSTM, the vanilla transformer does not process the input sequences serially but enters the entire sequence at once, which does not reflect the sequence information of the input sequence. Therefore, positional encoding plays the role of reflecting the sequence information in the input sequence. In this case, the relative sequence information of the input sequences is extracted as a vector with the same dimension as the input vector. The extracted positional encoding output and input are added element-wise to assign the sequence information to the input. At this point, the transformer uses periodic functions, such as sine and cosine, to efficiently assign the sequence positional information.

#### 2) MULTI-HEAD ATTENTION [38], [41]

The researchers who proposed the transformer also introduced the concept of multi-head attention to speed up the computation by parallel processing. The multi-head attention is a structure comprising multiple scaled dot-product attentions, where the number of attentions is represented by the number of heads. The reference formula for the multi-head attention on an input sequence  $X$  is shown in (1), and the specific process is as follows: the attention dimension is divided by the number of heads, and attention is performed individually. The extracted attention values are concatenated and multiplied by the output weight matrix ( $W^O$ ) to obtain the final attention value.

The attention performed by each head is self-attention and it is used to determine how each point or word in a

sequence is related to each other; it is expressed as a dictionary and takes as input three vectors: query (Q), key (K), and value (V). The AM applied by the transformer uses scaled dot-product attention to determine the weights of the query and the key. Through this process, the attention value, which is the weighted vector, is obtained. The formula referred to by the vanilla transformer is shown in (2), and the square root of K dimensions ( $\sqrt{d_k}$ ) is the scaling factor of scaled dot-product attention.

The output of the multi-head attention has the same dimensionality as the input, which makes it possible to stack multiple transformer blocks. Furthermore, stacked blocks have the same structure but do not share the same parameters.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V = \text{head}_i \quad (1)$$

$$\text{Attention}_h(X) = \text{concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (2)$$

### 3) POSITION-WISE FEED-FORWARD NEURAL NETWORK [38], [41]

The position-wise feed-forward neural network (FFNN) comprises FFNN and rectified linear unit (ReLU). By applying the ReLU function, nonlinearity is added to the model to train complex patterns more efficiently. Furthermore, the ReLU function is applied independently to each position input within the same encoder block so that different features can be extracted and trained from each position, as shown in (3).

$$\text{FFNN}(X_{AM}) = \max(0, X_{AM}W_1 + b_1)W_2 + b_2 \quad (3)$$

### 4) RESIDUAL CONNECTION AND LAYER NORMALIZATION [38], [41]

Residual connection is a concept that emerged from residual neural networks (ResNet). It solves the problem that the slope of the backpropagation process decreases as the depth of the model increases, preventing it from training. Therefore, the transformer applies residual connections in the internal modules of the encoder and decoder blocks to avoid the loss of input features. Furthermore, it applied layer normalization, an effective normalization method for time series models, to improve generalization performance. Additionally, both concepts were applied to multi-head attention and position-wise FFNN using (4) and (5), respectively.

$$X_{AM} = \text{LayerNorm}(\text{Attention}_h(X) + X) \quad (4)$$

$$X_{FFNN} = \text{LayerNorm}(X_{AM} + \text{FFNN}(X_{AM})) \quad (5)$$

## B. PROPOSED MODEL FOR FAULT DETECTION OF HYDRAULIC PUMPS

The vanilla transformer is used for translation to perform many-to-many tasks. However, detecting the fault of a hydraulic pump for an excavator is a many-to-one task, so the model cannot be used as it is. Therefore, we modified the inner structure of the vanilla transformer to achieve the purpose of this study. Roughly, the decoder and embedding layer

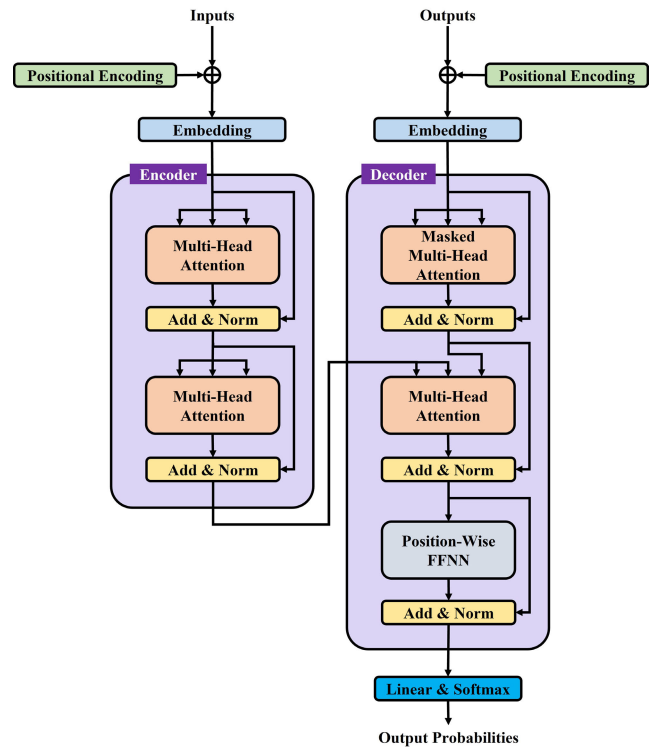


FIGURE 3. Architecture of the vanilla transformer [40].

of the vanilla transformer were deleted, and global average pooling and dropout in the classification part were added. The modified structure is presented in Fig. 4 and is largely divided into input, feature extraction, and classification. The detailed explanation is as follows:

First, the part of input directly uses 1D pump outlet pressure data without any conversion as input data. Afterward, the  $1 \times$  data points vector reflecting the relative position of the input is generated through positional encoding. Then, the input vector containing position information of the sequence is extracted through element-wise addition of the input data and output vector of positional encoding. At this time, the proposed model does not use an embedding layer. The reason for the above is that the proposed method uses 1D signal data as input, unlike the vanilla transformer that receives sentences as input, so there is no need for vectorization assumptions. Additionally, unlike language, the proposed model operates only when pressure data of a certain length are input which has no practical meaning such as special characters. So, we also removed the padding-masked module. Here, the padding masked module is a device that prevents focus on the word by inserting a very small negative value when paying attention to a word that has no actual meaning in a sentence.

Second, the encoder block structure of the vanilla transformer is used without transformation in the part of feature extraction. At this time, the number and dimensions of the encoder block are selected through optimization as shown in Section IV-B to be suitable for detecting multiple faults of hydraulic pumps.

Finally, we use a 1D global average pooling layer to transform the output values of the encoder into the input format of a fully connected (FC) layer in the part of the classification. The reason of that the proposed model used for classification problems outputs probabilities, unlike the existing vanilla transformer that outputs words. Therefore, the dimensionality reduction process is performed as above. Subsequently, we use the ReLU activation function which has the advantages of a fast convergence rate, high computational efficiency, and prevention of gradient vanishing. Additionally, dropout is applied to prevent overfitting, and then the softmax function is added to perform multi-fault detection. The softmax function plays a role in the activation function used in the output layer to classify three or more classes. The loss function also considers categorical cross-entropy, which is used in multi-class classification problems.

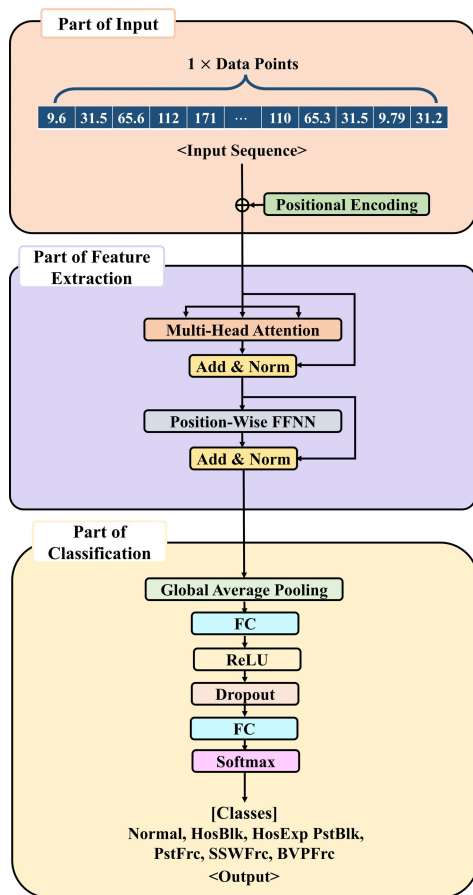


FIGURE 4. The architecture of the modified transformer.

#### IV. ALGORITHM OPTIMIZATION

This section describes the process of optimizing the models and presents the optimization results. Models are written in Python 3.8 with TensorFlow and run on Ubuntu 20.04 with the INTEL i7-13700K CPU, 64GB memory storage space, and NVIDIA GeForce RTX 4080 VGA.

To design models with high performance and short running time of an algorithm, LSTM and BiLSTM consider

monolayers, and the transformer considers parameters smaller than those proposed by the vanilla transformer. At this time, the optimizer uses adaptive moment estimation (Adam) and the learning rate is set to 0.001 which is considered appropriate in Adam.

Time series data have a long-term increasing or decreasing trend which may not be included in short data. Therefore, the optimization is performed with 2,000 data points by the grid search technique in the order of structure and hyperparameters. The information on the parameters considered for optimization is listed in Table 3 and Table 4. Grid search is used in most studies because it is easy to implement and allows the analysis of desired combinations. An epoch of 50 and a batch size of 16 were considered when performing the structural optimization.

Moreover, it is important to properly evaluate the model performance during the optimization process. Performance indicators such as accuracy, precision, recall, and F1-score are mainly measured to evaluate a classification model. These values are calculated based on the confusion matrix that is for comparing predicted values and actual values. Among these, accuracy is the most intuitive and simple classification performance indicator and it is suitable for use when the number of data per class is uniform. Accuracy is located on the diagonal of the confusion matrix and it is the proportion of the number of correctly predicted data in the number of total data. Precision is the proportion of data that are predicted correctly in data predicted to be true. Recall is the ratio of what the model predicts to be true compared to the fact. There is a trade-off between precision and recall described above. Therefore, the F1-score which is defined as the harmonic mean of precision and recall is also widely used. In particular, the F1-score is appropriate as a classification performance indicator when the number of data per class is not uniform [42], [43], [44]. This study selects accuracy as a classification performance indicator because the number of data per class is uniform. This indicator has a value between 0 and 100%, a value close to 100% means that be well classified.

If the model structure and hyperparameter information are not perfectly optimized, it will not show stable performance. Whereas to use this method in actual fields, it is necessary to design a model that shows stable and high performance. Thus, for each case, we have been trained 10 times and extracted accuracy to ensure the reliability of the results of optimization. Afterward, the average value of the 10 extraction accuracies is used as a classification performance indicator and called average accuracy. Moreover, for hydraulic pump fault detection, a model that performs well with or without noise should be selected. Thus, during the optimization process, the average accuracies without noise and with high noise were averaged and used as the criteria for parameter selection. Here, we added Gaussian noise with zero mean and standard deviation of 7 to the test dataset to simulate the high-noise scenario. Fig. 5 shows the steady-state data based on the noise level, and it can be seen that a standard deviation of 7 is a severe level.

**TABLE 3. Variations on the comparison models.**

Model	Epoch	Batch Size	Hidden Unit	DO 1 <sup>1</sup>	Attention Unit	DO 2 <sup>2</sup>
LSTM, BiLSTM	25	16	32	0	X	X
			64	0.2		
			128	0.3		
			256	0.5		
LSTM AM, BiLSTM AM	100	32	X	X	32	0
					64	0.2
					128	0.3
					256	0.5

<sup>1</sup> Dropout rate behind hidden unit.<sup>2</sup> Dropout rate behind attention unit.**TABLE 4. Variations on the modified transformer.**

Model	Epoch	Batch Size	Number of Heads	Encoder Layer	d <sub>model</sub> <sup>1</sup>	d <sub>ff</sub> <sup>2</sup>	Dropout Rate
Modified Transformer	25	16	2	3	32	32	0
	50				64	64	0.2
	75				128	128	0.3
	100				256	256	0.5

<sup>1</sup> Dimension of model.<sup>2</sup> Dimension of feed-forward.**TABLE 5. Results of structure optimization for LSTM and BiLSTM.**

Model	Hidden Unit	Dropout Rate	Average Accuracy (%)		
			Std 0 <sup>1</sup>	Std 7 <sup>2</sup>	Avg <sup>3</sup>
LSTM	256	0.5	96.49	57.83	77.16
BiLSTM	256	0.2	97.98	57.15	77.57

<sup>1</sup> Standard deviation 0.<sup>2</sup> Standard deviation 7.<sup>3</sup> Average of standard deviation 0 and standard deviation 7.**TABLE 6. Results of structure optimization for AM combined models.**

Model	Attention Unit	Dropout Rate	Average Accuracy (%)		
			Std 0 <sup>1</sup>	Std 7 <sup>2</sup>	Avg <sup>3</sup>
LSTM AM	64	0	93.18	48.99	71.09
BiLSTM AM	32	0	97.90	48.36	73.13

<sup>1</sup> Standard deviation 0.<sup>2</sup> Standard deviation 7.<sup>3</sup> Average of standard deviation 0 and standard deviation 7.

## A. COMPARISON MODELS

First, we optimized the structure of the comparison groups, which are the LSTM family models such as LSTM, BiLSTM, LSTM AM, and BiLSTM AM. The optimization results after comparing the performance of LSTM and BiLSTM according to the hidden unit and dropout rate are shown in Table 5. Here, the optimized LSTM and BiLSTM structures are used to optimize the LSTM AM and BiLSTM AM structures. We considered Bahdanau AM because of

its excellent performance in a previous study [45]. Table 6 shows the optimization results and compares the performance according to the dropout rate applied to the attention unit and AM. Table 5 and Table 6 show that after performing the structural optimization, BiLSTM performs marginally better than LSTM with an average accuracy difference of less than 1%. However, after applying AM, the average accuracy of BiLSTM AM is 2.05% higher than that of LSTM AM.



**TABLE 7. Results of hyperparameter optimization for comparison models.**

Model	Epoch	Batch Size	Hidden Unit	DO 1 <sup>1</sup>	Attention Unit	DO 2 <sup>2</sup>	Average Accuracy (%)		
							Std 0 <sup>3</sup>	Std 7 <sup>4</sup>	Avg <sup>5</sup>
LSTM	50	32	256	0.5	X	X	98.68	59.66	<b>79.17</b>
BiLSTM	25	16	256	0.2	X	X	97.11	58.39	77.75
LSTM AM	25	32	256	0.5	64	0	95.40	50.36	72.88
BiLSTM AM	75	32	256	0.2	32	0	97.32	50.49	73.91

<sup>1</sup> Dropout rate behind hidden unit.<sup>2</sup> Dropout rate behind attention unit.<sup>3</sup> Standard deviation 0.<sup>4</sup> Standard deviation 7.<sup>5</sup> Average of standard deviation 0 and standard deviation 7.**TABLE 8. Results of structure optimization for modified transformer 1.**

Encoder Block	d <sub>model</sub>	d <sub>ff</sub>	Dropout Rate	Average Accuracy (%)		
				Std 0 <sup>1</sup>	Std 7 <sup>2</sup>	Avg <sup>3</sup>
3	128	128	0	100.00	56.26	78.13
6	64	256	0	100.00	58.06	79.03

<sup>1</sup> Standard deviation 0.<sup>2</sup> Standard deviation 7.<sup>3</sup> Average of standard deviation 0 and standard deviation 7.**TABLE 9. Results of structure optimization for modified transformer 2.**

Encoder Block	Number of Heads	Average Accuracy (%)		
		Std 0 <sup>1</sup>	Std 7 <sup>2</sup>	Avg <sup>3</sup>
3	2	100.00	56.26	78.13
	4	100.00	56.56	78.28
	8	100.00	59.49	79.75
6	2	100.00	58.06	79.03
	4	100.00	53.35	76.68
	8	100.00	55.97	77.99

<sup>1</sup> Standard deviation 0.<sup>2</sup> Standard deviation 7.<sup>3</sup> Average of standard deviation 0 and standard deviation 7.**TABLE 10. Results of hyperparameter optimization for modified transformer.**

Model	Epoch	Batch Size	Encoder Block	Number of Heads	d <sub>model</sub>	d <sub>ff</sub>	Dropout Rate	Average Accuracy (%)		
								Std 0 <sup>1</sup>	Std 7 <sup>2</sup>	Avg <sup>3</sup>
Modified Transformer	50	16	3	8	128	128	0	100.00	59.49	79.75

<sup>1</sup> Standard deviation 0.<sup>2</sup> Standard deviation 7.<sup>3</sup> Average of standard deviation 0 and standard deviation 7.

Next, we performed hyperparameter optimization. The optimization results for the comparison group are shown in Table 7. The performance of the optimized models shows that

the lightest model, LSTM, exhibits the highest performance of 79.17%. Moreover, LSTM and BiLSTM perform worse when AM is applied.

## B. PROPOSED MODEL

First, we compared the performance according to  $d_{\text{model}}$  (model dimension),  $d_{\text{ff}}$  (feedforward dimension), and dropout rate, respectively when the encoder block is considered to be 3 and 6. The results are shown in Table 8. This table shows that the difference between them in average accuracy is less than 1% although the encoder block 6 outperforms 3. Therefore, we analyzed the number of heads in encoder blocks 3 and 6 (Table 9). As a result, it shows substantially large differences in performance depending on the number of heads. Next, we optimized the hyperparameters of the proposed model based on the optimized encoder block and number of heads. The results are shown in Table 10.

## V. SIMULATION AND RESULT

Hydraulic pumps have vibration from the machine itself, and pumps used in excavators in particular present a lot of noise coming from outside. Therefore, we must use a noise-robust model for fault detection of hydraulic pumps. In this section, we compared the performance and running time of the models, and then examined how they are affected by data points and noise.

### A. PERFORMANCE OF MODELS

The performance results of the optimized models are shown in Table 11 when additional noise is absent. The modified transformer performs the best with 100%. Moreover, the operating time of the algorithm is the shortest after LSTM and LSTM AM at 0.00271 s. Comparing the modified transformer with the two models, there is a very slight difference of 0.00036 s and 0.00009 s respectively. Based on the above results, we assess that the modified transformer will enable fast and accurate fault detection. Afterward, we extract the performance of the models in the form of a confusion matrix as shown in Fig. 5 to analyze which faults were misclassified. The figure shows that the modified transformer classifies all faults, whereas the other models misclassify the normal and three faults caused by friction, as predicted in Section II-B. These results show that the modified transformer has high accuracy for faults that are difficult to classify due to the high similarity of signal patterns.

### B. PERFORMANCE ACCORDING TO DATA POINTS

We visualize in a chart to confirm the changing trend according to data points as shown in Fig. 6. The figure confirms that all models show a high fault detection accuracy of greater than or equal to 95.40% for the 2,000 data points for which optimization was performed. Among them, only the modified transformer shows a fault detection accuracy of 100%. Moreover, unlike the modified transformer, which showed a high performance of 99.93% for all data points, the other comparison models have significantly decreased performance at data points (1,000 and 1,500) for which no optimization was performed. But these models show high performance of greater than or equal to 95.38% in a relatively

short 500 data points. The reason is that time series models do not cause long-term dependency problems when the number of data points is small (500). However, comparison models do not solve the long-term dependency problem when the number of data points is large (1,000 and 1,500). Based on the above results, it can be said that the modified transformer has the highest and most stable performance compared to other models even if the data point changes. Therefore, the modified transformer can be used for various fault detection.

**TABLE 11. Comparison of accuracy and algorithm operation time between optimized models.**

Model	Test Time (s)	Average Accuracy (%)
LSTM	0.00235	98.68
BiLSTM	0.00434	97.11
LSTM AM	0.00262	95.40
BiLSTM AM	0.00470	97.32
Modified Transformer	0.00271	100.00

### C. PERFORMANCE ACCORDING TO NOISE LEVELS

We confirm the performance of each model according to additional noise, and the data points are set to 2,000, which is the number used for optimization. The additional noise is considered Gaussian noise with zero mean.

First, we change the standard deviation to check the effect according to noise level as shown in Fig. 7 with dynamic pressure data of Normal. At this time, the standard deviation is considered to be 0 to 9. As shown in this figure, we find that the characteristics of data are maintained up to standard deviation 4 whereas the characteristics of data start to disappear from above standard deviation 5. At standard deviation 6, a lot of extra noise comes in and the characteristics of data largely disappear, but it is still possible to classify with the naked eye. Furthermore, it can be seen that the noise level is so high that is difficult to distinguish with the naked eye when the standard deviation is 7 or more. Next, we conduct a similarity test using SSIM as in Section II to quantify the effect of the noise level. The degree of change according to the standard deviation is confirmed based on the noise-free (standard deviation 0), and the results are shown in Table 12. At this time, the value is seen that becomes less than or equal to 0.9 and the similarity decreases significantly when the standard deviation is 6 or more. Also, the similarity result is less than or equal to 85% when the standard deviation is 8 or more.

As shown in Fig. 7 and Table 12, we find that the deformation of a signal is large above standard deviation 6 and it is difficult to determine the same signal data because the deformation of a signal is too large above standard deviation 8. Thus, we confirm a performance up to a standard deviation of 7 that the similarity test result is greater than or equal to 85% to check performance when severe external noise is coming while maintaining data characteristics in this study.

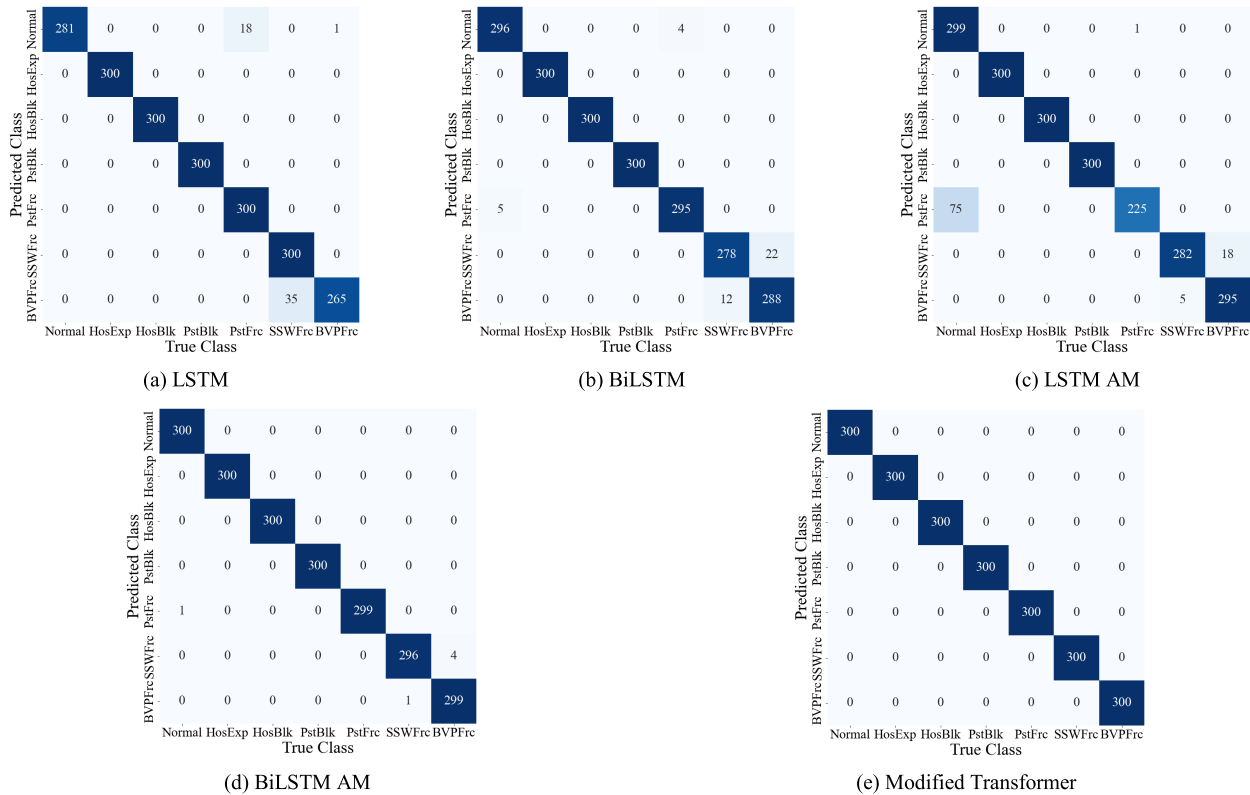


FIGURE 5. Confusion matrices of the test dataset in the first trial.

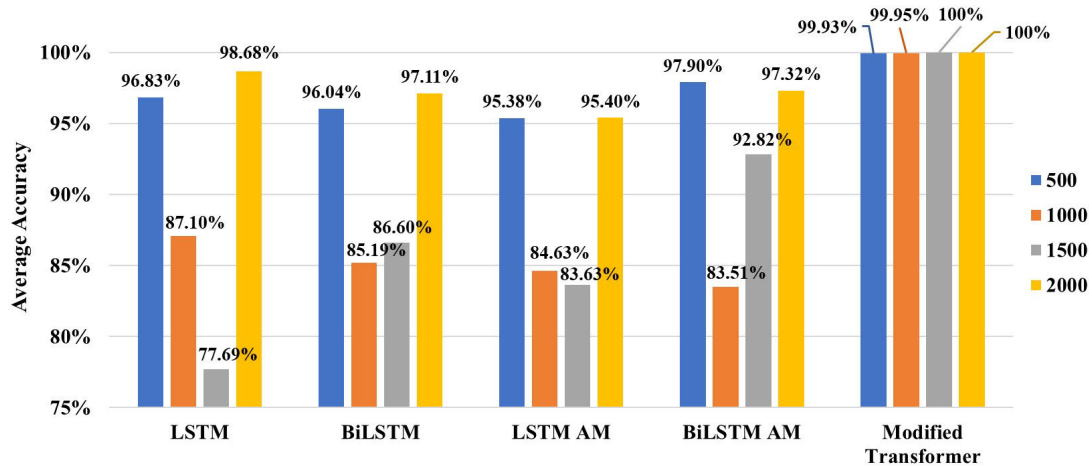


FIGURE 6. Average accuracy according to data points.

TABLE 12. Similarity results according to standard deviation.

	Standard Deviation (%)								
	1	2	3	4	5	6	7	8	9
SSIM	0.9964	0.9857	0.9677	0.9481	0.9232	0.8895	0.8611	0.8112	0.7696

The performance results for each model according to the additional noise are shown in Table 13, and we visualize it in a chart as shown in Fig. 8 to check the changing trend.

As shown in this table, the modified transformer shows a high fault detection accuracy of greater than or equal to 98% up to a standard deviation 4, unlike fault detection accuracy of

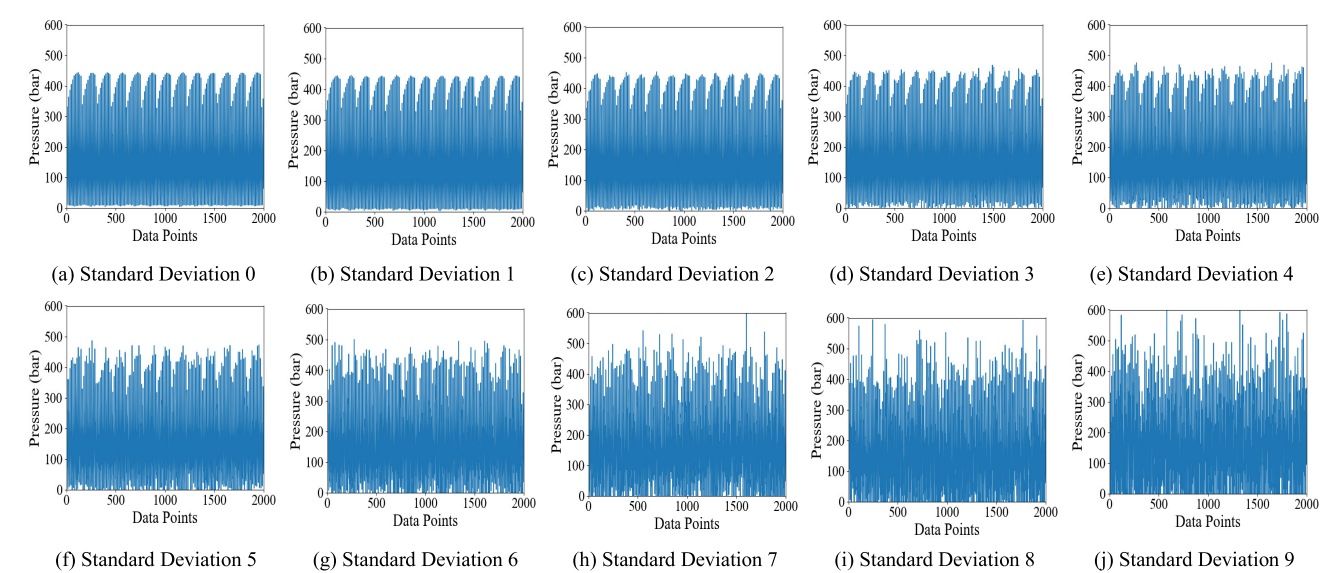


FIGURE 7. Dynamic pressure data of hydraulic pump according to noise levels.

TABLE 13. Average accuracy according to models in 2,000 data points.

Model	Standard Deviation (%)							
	0	1	2	3	4	5	6	7
LSTM	98.68	98.61	98.14	95.22	88.07	78.41	69.08	<b>59.66</b>
BiLSTM	97.11	97.13	95.97	91.92	83.27	74.38	66.23	58.39
LSTM AM	95.40	95.34	94.39	90.53	82.03	71.42	61.82	50.36
BiLSTM AM	97.32	97.26	96.70	93.41	82.80	70.42	60.14	50.49
Modified Transformer	<b>100.00</b>	<b>100.00</b>	<b>100.00</b>	<b>99.93</b>	<b>98.08</b>	<b>85.44</b>	<b>70.63</b>	59.49

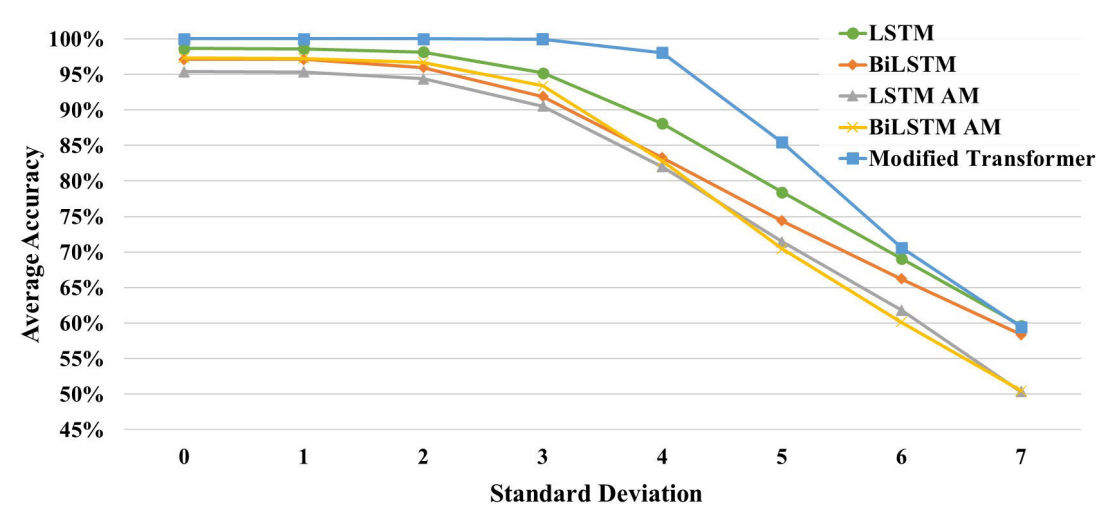


FIGURE 8. Visualization of average accuracy according to noise level in 2,000 data points.

comparison models falls less than or equal to 90%. However, the modified transformer shows a fault detection accuracy of less than or equal to 85.44% from above standard deviation 5 that has begun to disappear data characteristics, although it

shows higher performance than the comparison group which showed less than 78.41% performance. Based on the above results, the modified transformer has a high performance up to an additional noise level at which the characteristics of the training data are maintained, unlike comparison models. However, the performance decreases significantly at noise level (standard deviation 6 and 7) which largely disappeared data characteristics due to severe noise and the classification performance is similar to comparison models.

## VI. CONCLUSION

In this study, we performed modified transformer-based fault detection of hydraulic pumps using pressure signal data. For the fault detection of hydraulic pumps, we modified the transformer, which performs excellently in the translation task. Further, we proposed the modified transformer, a lightweight model, which is optimized with fewer parameters than the vanilla transformer. The modified transformer has a fault detection time with an algorithm operation time of 0.000271 s, which is almost the same as the single LSTM with the shortest operation time. Additionally, the proposed model showed 100% accurate fault detection accuracy, whereas other time series models misclassified Normal and friction-induced faults of similar shape and size. After that, we performed an analysis of the models according to the number of data points. At this time, only the modified transformer showed stable and high performance at all data points for which no optimization was performed (500, 1000, and 1500). Next, performance analysis for each model was performed according to noise level because sensor data reflected various noises collected in actual industrial sites. The modified transformer showed the best performance of 100% when no additional noise is present and a high performance of greater than or equal to 98.08% up to a Gaussian noise with zero mean and standard deviation 4 where the data characteristics are preserved. However, the performance of the proposed model decreased to less than or equal to 90% since the standard deviation 5 where data characteristics start to disappear. Also, we confirmed that pressure signals could be used for fault diagnosis rather than commonly used vibration signals.

Therefore, if noise is absent or little, we found that the modified transformer has stable and high performance when compared to other time series models. However, the performance decreases significantly and is not different from comparison models when the noise is large (standard deviation 6 and 7). Furthermore, the obtained data based on real sensors contain missing values, but this paper did not consider that. Therefore, in the future, we plan to design a modified transformer-based fault detection model to have stable and excellent fault detection performance even if additional noise is large or missing values occur.

## ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their feedback.

## REFERENCES

- [1] S. Tang, S. Yuan, and Y. Zhu, "Deep learning-based intelligent fault diagnosis methods toward rotating machinery," *IEEE Access*, vol. 8, pp. 9335–9346, 2020.
- [2] C. Junsheng, Y. Dejie, and Y. Yu, "A fault diagnosis approach for roller bearings based on EMD method and AR model," *Mech. Syst. Signal Process.*, vol. 20, no. 2, pp. 350–362, Feb. 2006.
- [3] Y. Xu, Y. Sun, J. Wan, X. Liu, and Z. Song, "Industrial big data for fault diagnosis: Taxonomy, review, and applications," *IEEE Access*, vol. 5, pp. 17368–17380, 2017.
- [4] Y. Yang, L. Ding, J. Xiao, G. Fang, and J. Li, "Current status and applications for hydraulic pump fault diagnosis: A review," *Sensors*, vol. 22, no. 24, p. 9714, Dec. 2022.
- [5] P. Casoli, M. Pastori, F. Scolari, and M. Rundo, "A vibration signal-based method for fault identification and classification in hydraulic axial piston pumps," *Energies*, vol. 12, no. 5, p. 953, Mar. 2019.
- [6] Y. Tian, C. Lu, and Z. L. Wang, "Approach for hydraulic pump fault diagnosis based on WPT-SVD and SVM," *Appl. Mech. Mater.*, vols. 764–765, pp. 191–197, May 2015.
- [7] S. Tang, S. Yuan, Y. Zhu, and G. Li, "An integrated deep learning method towards fault diagnosis of hydraulic axial piston pump," *Sensors*, vol. 20, no. 22, p. 6576, Nov. 2020.
- [8] Y. Zhu, G. Li, R. Wang, S. Tang, H. Su, and K. Cao, "Intelligent fault diagnosis of hydraulic piston pump based on wavelet analysis and improved AlexNet," *Sensors*, vol. 21, no. 2, p. 549, Jan. 2021.
- [9] Y. Zhu, G. Li, R. Wang, S. Tang, H. Su, and K. Cao, "Intelligent fault diagnosis of hydraulic piston pump combining improved LeNet-5 and PSO hyperparameter optimization," *Appl. Acoust.*, vol. 183, Dec. 2021, Art. no. 108336.
- [10] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5990–5998, Jul. 2018.
- [11] Y. Zhu, T. Zhou, S. Tang, and S. Yuan, "A data-driven diagnosis scheme based on deep learning toward fault identification of the hydraulic piston pump," *J. Mar. Sci. Eng.*, vol. 11, no. 7, p. 1273, Jun. 2023.
- [12] H. Tao, P. Jia, X. Wang, and L. Wang, "Real-time fault diagnosis for hydraulic system based on multi-sensor convolutional neural network," *Sensors*, vol. 24, no. 2, p. 353, Jan. 2024.
- [13] S. Tang, Y. Zhu, and S. Yuan, "An adaptive deep learning model towards fault diagnosis of hydraulic piston pump using pressure signal," *Eng. Failure Anal.*, vol. 138, Aug. 2022, Art. no. 106300.
- [14] S. Tang, Y. Zhu, and S. Yuan, "Intelligent fault identification of hydraulic pump using deep adaptive normalized CNN and synchrosqueezed wavelet transform," *Rel. Eng. Syst. Saf.*, vol. 224, Aug. 2022, Art. no. 108560.
- [15] V. Muralidharan, V. Sugumaran, and V. Indira, "Fault diagnosis of monoblock centrifugal pump using SVM," *Eng. Sci. Technol., Int. J.*, vol. 17, no. 3, pp. 152–157, Sep. 2014.
- [16] A. Kumar, C. P. Gandhi, Y. Zhou, R. Kumar, and J. Xiang, "Improved deep convolution neural network (CNN) for the identification of defects in the centrifugal pump using acoustic images," *Appl. Acoust.*, vol. 167, Oct. 2020, Art. no. 107399.
- [17] D. Xiao, Y. Huang, X. Zhang, H. Shi, C. Liu, and Y. Li, "Fault diagnosis of asynchronous motors based on LSTM neural network," in *Proc. Prognostics Syst. Health Manage. Conf. (PHM-Chongqing)*, Chongqing, China, Oct. 2018, pp. 540–545.
- [18] L. Yu, J. Qu, F. Gao, and Y. Tian, "A novel hierarchical algorithm for bearing fault diagnosis based on stacked LSTM," *Shock Vib.*, vol. 2019, no. 1, pp. 1–10, Jan. 2019.
- [19] K. Shen and D. Zhao, "An EMD-LSTM deep learning method for aircraft hydraulic system fault diagnosis under different environmental noises," *Aerospace*, vol. 10, no. 1, p. 55, Jan. 2023.
- [20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth  $16 \times 16$  words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [21] B. Wu, W. Cai, F. Cheng, and H. Chen, "Simultaneous-fault diagnosis considering time series with a deep learning transformer architecture for air handling units," *Energy Buildings*, vol. 257, Feb. 2022, Art. no. 111608.
- [22] J. Li, Y. Bao, W. Liu, P. Ji, L. Wang, and Z. Wang, "Twins transformer: Cross-attention based two-branch transformer network for rotating bearing fault diagnosis," *Measurement*, vol. 223, Dec. 2023, Art. no. 113687.



- [23] D. Wang, J. Zhang, S. Liu, F. Lyu, W. Huang, and B. Xu, "KD-ViT: A lightweight method for online wear state recognition of key friction pairs in axial piston pump," *IEEE Trans. Ind. Informat.*, vol. 20, no. 7, pp. 9621–9632, Jul. 2024.
- [24] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proc. AAAI Conf. Artif. Intell.*, Washington, DC, USA, 2023, pp. 11121–11128.
- [25] Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov, Y. Wang, and J. Ren, "EfficientFormer: Vision transformers at MobileNet speed," 2022, *arXiv:2206.01191*.
- [26] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: A compact task-agnostic BERT for resource-limited devices," 2020, *arXiv:2004.02984*.
- [27] A. Esposito, *Fluid Power With Application*, 7th ed., Englewood Cliffs, NJ, USA: Prentice-Hall, 2009.
- [28] A. R. Kim, H. S. Kim, H. Y. Ra, H. S. Jeong, and S. Y. Kim, "Performance comparison of LSTM-based networks for fault detection in hydraulic pump system," in *Proc. KSME Spring Conf.*, Muju, South Korea, 2023, p. 42.
- [29] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [30] B. Liu, Z. Zhang, and R. Cui, "Efficient time series augmentation methods," in *Proc. 13th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Chengdu, China, Oct. 2020, pp. 1004–1009.
- [31] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, and H. Xu, "Time series data augmentation for deep learning: A survey," 2020, *arXiv:2002.12478*.
- [32] G. Iglesias, E. Talavera, Á. González-Prieto, A. Mozo, and S. Gómez-Canaval, "Data augmentation techniques in time series domain: A survey and taxonomy," *Neural Comput. Appl.*, vol. 35, no. 14, pp. 10123–10145, Mar. 2023.
- [33] B. K. Iwana and S. Uchida, "An empirical survey of data augmentation for time series classification with neural networks," *PLoS ONE*, vol. 16, no. 7, Jul. 2021, Art. no. e0254841.
- [34] C. Yang, Z. Qiao, R. Zhu, X. Xu, Z. Lai, and S. Zhou, "An intelligent fault diagnosis method enhanced by noise injection for machinery," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.
- [35] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," *Inst. Cogn. Sci., Univ. California, San Diego, La Jolla, CA, USA*, Tech. Rep. ADA164453, 1985.
- [36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [37] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*.
- [38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long beach, CA, USA, 2017, pp. 5998–6008.
- [39] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 3104–3112.
- [40] M. Petrolini, S. Cagnoni, and M. Mordonini, "Automatic detection of sensitive data using transformer-based classifiers," *Future Internet*, vol. 14, no. 8, p. 228, Jul. 2022.
- [41] Y. Ding, M. Jia, Q. Miao, and Y. Cao, "A novel time–frequency transformer based on self-attention mechanism and its application in fault diagnosis of rolling bearings," *Mech. Syst. Signal Process.*, vol. 168, Apr. 2022, Art. no. 108616.
- [42] M. Grandini, E. Bagli, and G. Visani, "Metrics for multi-class classification: An overview," 2020, *arXiv:2008.05756*.
- [43] M. A. Razib, D. Javeed, M. T. Khan, R. Alkanhel, and M. S. A. Muthanna, "Cyber threats detection in smart environments using SDN-enabled DNN-LSTM hybrid framework," *IEEE Access*, vol. 10, pp. 53015–53026, 2022.
- [44] S. Asif, W. Yi, Q. U. Ain, J. Hou, T. Yi, and J. Si, "Improving effectiveness of different deep transfer learning-based models for detecting brain tumors from MR images," *IEEE Access*, vol. 10, pp. 34716–34730, 2022.
- [45] A. R. Kim, H. S. Kim, H. Y. Ra, and S. Y. Kim, "Attention-based fault detection in hydraulic pump system," in *Proc. KSME Conf.*, Incheon, South Korea, 2023, pp. 858–860.



**A. RAN KIM** received the B.S. degree in mechanical convergence system engineering and the M.S. degree in mechanical engineering from Kunsan National University, Gunsan, Republic of Korea, in 2021 and 2024, respectively. From 2021 to 2022, she was a Research Assistant with the Engineering Research Center, Kunsan National University. Her research interests include navigation systems, autonomous driving, deep learning, target detection and classification, and SLAM.



**HA SEON KIM** received the B.S. degree in mechanical convergence system engineering and the M.S. degree in mechanical engineering from Kunsan National University, Gunsan, Republic of Korea, in 2022 and 2024, respectively. His research interests include navigation systems, autonomous driving, deep learning, target detection and classification, and SLAM.



**SUN YOUNG KIM** (Member, IEEE) received the B.S. degree in electronic engineering from Kookmin University, in 1999, and the M.S. and Ph.D. degrees in mechanical and aerospace engineering from Seoul National University, in 2015 and 2019, respectively. From 2019 to 2020, she was a Postdoctoral Researcher with the School of Intelligent Mechatronics Engineering, Sejong University, Seoul, Republic of Korea. From 2020 to 2023, she was an Assistant Professor with the School of Mechanical Engineering, Kunsan National University, Jeonbuk State, Republic of Korea, where she has been an Associate Professor, since 2023. Her research interests include navigation systems, filtering, localization, GNSS interference detection and mitigation, multi-sensor fusion, multi-target tracking, target detection and classification, SLAM, autonomous vehicles, and deep learning.

...