Pattern Recognition 172 (2026) 112497

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

A novel vision transformer with selective residual in multihead self-attention for pattern recognition

Arun Kumar Sharma*, Nishchal K. Verma

Dept. of Electrical Engineering, Indian Institute of Technology, Kanpur, 208016, U.P., India

ARTICLE INFO

Keywords:
Intelligent fault diagnosis
Pattern recognition
Spectrum transforms of raw signal
Multi-head attention
Accumulative attention
Manhattan norms
Short-term fourier transform

ABSTRACT

Intelligent fault diagnosis requires robust capturing of specific features, representing the fault patterns, from time-series vibration signals. Most of the existing solutions require complex preprocessing steps to make the signal suitable for training a deep learning model. This article presents a novel vision transformer with a selective residual in the multihead self-attention network, called Selective Residual Vision Transformer (SeReViT), for improved robustness in capturing the fault signature from the vibration signal. The novel attention mechanism incorporates cumulative attention by utilizing the best attention through residual connections in each block of multihead attention. The best attention term is defined using the highest value of L1-norms of attention value (the scaled-dot product of key and query) of multiheads. It enables the model to focus on selected best attention to learn the long-range dependencies among sequential input image patches, resulting in better classification performance. The proposed framework is validated for fault diagnosis on the Case Western Reserve University bearing fault diagnosis dataset and the Paderborn University dataset. Since these datasets are already cleaned data, noisy vibration data are created by adding white noise for the demonstration of the robustness of the proposed framework. The vibration signals are first converted to images using the short-time Fourier transform with a fixed window size. The generated images are used to train and validate the proposed SeReViT. The results outperformed the state-of-the-art convolution-based models for fault diagnosis for both cleaned datasets and noisy datasets. The short-time Fourier transform is utilized to convert the noisy (raw) vibration signals from rotating machines to spectrum images.

1. Introduction

With the advancements in modern computational technology, deep learning (DL)-based intelligent fault diagnosis has gained much attention from various researchers [1,2]. DL-based methods have shown remarkable performance in capturing complex patterns for image classification. Intelligent fault diagnosis requires accurate capturing of anomalies in the vibration signal acquired by suitable sensors in the running conditions of the machine [3–5]. The vibration signals recorded as time-series data also contain noise, which makes anomaly detection complex and inaccurate. Several methods, like domain-adversarial training of neural networks [6,7] and cross-domain fault diagnosis [8,9] have been attempted to deal with accurate fault diagnosis. EvoN2N [10] and GS-EvoN2N [11] have shown commendable performance for the selection of the best architecture of DNN to ensure accurate fault diagnosis even with a low number of samples in the target domain, assuming that a source model, trained on sufficient source data, is available for model initialization. These methods assume that vibration signals are well processed to remove possible noise contamination during recording.

Many research works have reported the application of convolutional-based deep learning models with short-time Fourier transform (STFT) to transform the raw vibration signals into images for fault classification [12–16]. Yan et al. [12] applied the Hilbert transform to convert the time-series signal into an envelope spectrum and then fed it to DCNN for automatic extraction of fault features. Cheng et al. [13] suggested the use of wavelet transform to extract time-frequency image features from raw time-series signals, followed by the use of a generative adversarial network to synthesize more images for data augmentation. The augmented data is used to train the CNN model. The diagnostic accuracy reported was higher even in a noisy working environment. Cheng et al. [14] suggested a continuous wavelet transform-local binary convolutional neural network (CWT-LBCNN) to avoid overfitting and faster training of the model and reported reliable performance compared to traditional convolutional-based models. Choudhary et al. [15]

* Corresponding author.
E-mail addresses: arunshr.iit@gmail.com, arnksh@iitk.ac.in (A.K. Sharma), nishchal@iitk.ac.in (N.K. Verma).

https://doi.org/10.1016/j.patcog.2025.112497
Received 5 December 2024; Received in revised form 17 August 2025; Accepted 23 September 2025
Available online 26 September 2025

introduced the application of non-invasive thermal imaging to train a LeNet-5-based convolutional model for fault identification of rotating machines. Liu et al. [16] suggested multi-source feature fusion followed by the application of the LightGBM model for fault classification. All these models rely on convolution-based feature extraction for fault classification. The convolutional-based methods have shown outstanding performance for fault diagnosis as well as various other computer vision applications. The process of convolution is capable of capturing only local spatial patterns using the wavelet-transformed envelope. For a diagnostic model to be capable of capturing global dependencies and contextual information in the image pattern, the attention mechanism has been proven to be a very effective solution.

Tiago et al. [17] have discussed the importance and effect of attention mechanisms for capturing contextual information from input images. Vision Transformer (ViT) by Alexey et al. [18] was the first to use the attention mechanism for image classification, inspired by transformer architecture for natural language processing (NLP). The vision transformer has been reported to outperform the CNN-based models if trained with a large number of samples [18]. In recent years, there have been various improvements in ViT, such as data-efficient image transformers & distillation through attention [19], hybridization with the CNN model [20], and swin transformer [21]. Several studies have been reported on the application of ViTs for image classification and analysis in different scenarios [22–26]. Wang et al. [22] introduced a multistage convolutional layer-based ViT-Plus model for the classification of the genitourinary syndrome of menopause using OCT images. They utilized the multi-stage Conv2d layer to convert the original images into patches rather than using a single Conv2d layer. Sabry et al. [23] introduced a hybrid mechanism using an auto-encoder, info-GAN, and vision transformer for image retrieval in an unsupervised manner. Rodrigo et al. [27] compare the performances of different variants of ViT with CNN-based models for face recognition applications. Diko et al. [24] introduced the application residual along with an attention module to enhance the feature diversity learning in the vision transformer, but it does not exploit the cumulative effect of the best-performing attention head. A recent work by Li et al. [25] demonstrated the performance of ViT for handwritten text recognition. Nie et al. [26] introduced scaleViT: a variant of the attention mechanism that exploits multiscale across spatial dimensions. Their results conclude that ViT outperformed CNN-based deep models in terms of accuracy and robustness against distance and occlusions for the aforementioned applications, even when the deepest CNN-based model is deployed.

The motivation behind this work is to investigate and develop an attention-based deep learning model capable of capturing global dependencies and contextual information from the STFT images of acoustic data of the bearings of rotating machines. These vibration data are always contaminated by various noises during their recording, usually of uniform distribution. Therefore, this study focuses on the development of a robust ViT model that has a cumulative effect to cancel out uniform noises in the multi-head mechanism. Different heads in the multi-head attention mechanism encode different relations among input tokens and produce different feature spaces using different subspaces [28]. Creating a stronger focus on the dominant head and adding a greater contextual confidence can make the model robust enough to counter the effect of noise in the aggregated feature representation. Therefore, we present a novel vision transformer with selective residual in multihead self-attention (SeReViT) for intelligent fault diagnosis. To the best of our knowledge, there is no study reported on the use of selective residual in multihead self-attention, where the residual skip connection is facilitated by the best of the attention heads. The key highlights of the contributions to this work are summarized below:

1) This work introduces a new multi-head attention mechanism using a residual connection of selective attention among multiple heads. The Manhattan norms, also known as the L1 norms, are applied to the attention scores from the multi-attention heads. The best of the L1 norms of the attention scores is used to select the attention head output as a residual add-on to the final projected attention output. This mechanism provides the multihead attention module to be biased towards the best learning head with an accumulative effect.
2) The proposed framework (SeReViT) is first pre-trained using a large number of samples obtained by applying STFT to raw time-series data from the Case Western Reserve University (CWRU) fault diagnostic data recorded at Drive End (DE) with the 7 mil fault diameter and zero motor load. The pre-trained Res-ViT is then fine-tuned using small samples from the target domain data: (i) CWRU with 21 mil fault diameters for different non-zero loads, and (iii) Paderborn University (PBU) data under load and operating conditions.

2. Theoretical background: Vision Transformer

The transformer introduced in the seminal paper “Attention is All You Need” by Vaswani et. al[29]. is a powerful architecture that revolutionized natural language processing (NLP) tasks. With the success of the transformer model in NLP, attention-based architecture was extended to a ground-breaking architecture called Vision Transformer (ViT). Introduced in the paper “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” by Dosovitskiy et al. in 2020 [18], it has been reported to achieve remarkable results in image classification tasks. Unlike traditional CNN models, which have been the dominant approach in computer vision, ViT relies solely on the self-attention mechanism of transformers to capture relationships between image patches.

The schematic architecture of the vision transformer and multi-head attention mechanism has been depicted in Fig. 1. Contrary to the transformer architecture, the vision transformer consists of only encoder blocks that accept the input image as a flattened sequence of

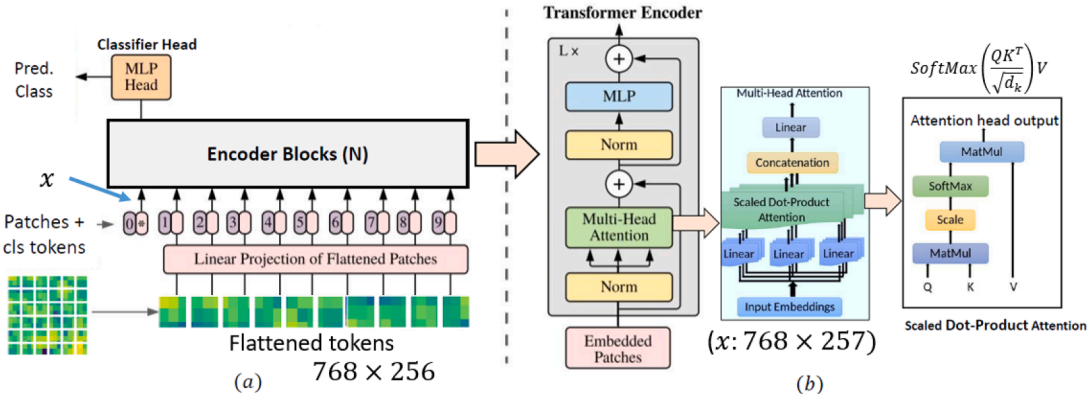

Fig. 1. Architecture of Vision Transformer for computer vision [18].

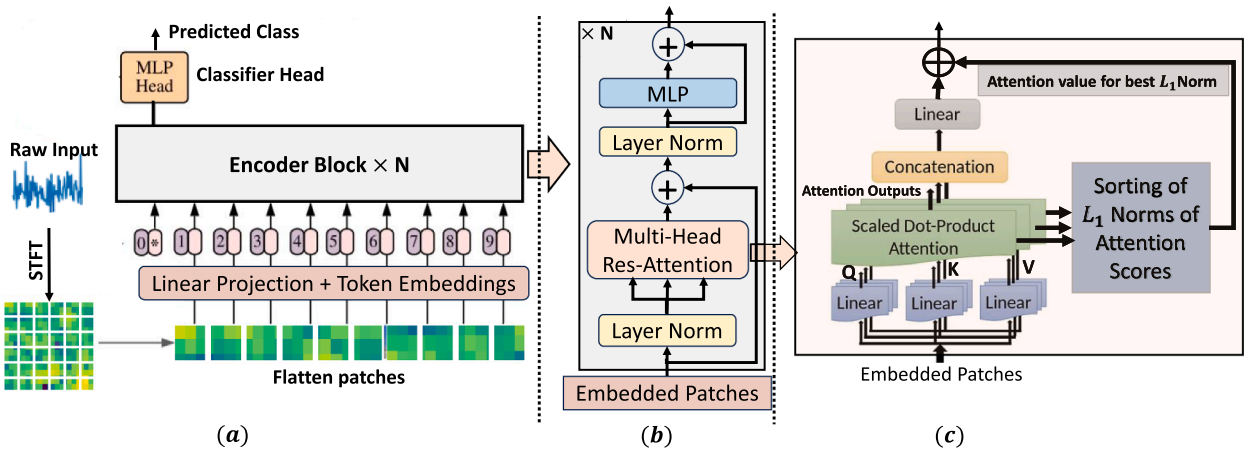

Fig. 2. The proposed selective residual vision transformer (SeReViT): (a) complete block diagram for transformation from raw input (vibration data) to output (predicted fault class); (b) unit block of the encoder with multi-head Res-Attention; and (c) detailed block diagram of the multi-head attention mechanism with residual connection for accumulative attention.

fixed-size patches. These patches are then embedded with class tokens to capture the class-wise visual information contained in the sequential patches. The image patches are then processed by the encoder block, which consists of multiple heads of self-attention and feed-forward neural networks (as shown in Fig. 1(b)).

3. SeReViT: Selective Residual Vision Transformer

This section presents the proposed SeReViT with selective residual in the multihead attention module. Since attention is the major component of ViT that captures the relationship between the sequential input patches, this section aims to present the new mechanism of multihead self-attention. Fig. 2 depicts the complete block diagram of the proposed framework.

The major blocks are explained below.

Input block: The raw vibration signal is first converted to spectral images using STFT (explained in Section 4.2) and then into spatial patches. Since the transformer architecture does not inherently possess knowledge of spatial relationships, positional embeddings are provided for the understanding of patch positions within the image. The learnable positional embeddings are provided to represent the spatial importance of each patch.

Multi-Head Attention The multi-head attention is a parallel combination of multiple self-attention heads. Each attention head receives input embeddings in the form of three components: Query (Q), Key (K), and Value (V) via three different linear layers with learnable parameters. The attention score of each head is computed using the softmax of the correlation between K and Q as shown in Eq. (2). The attention output of each head is defined in Eq. (3).

$$\mathbf{A}_c = \frac{QK^T}{\sqrt{d}} \tag{1}$$

$$\mathcal{A}_{score} = SoftMax\{\mathcal{A}_c\} \tag{2}$$

$$\mathcal{A} = V\mathcal{A}_{score} \tag{3}$$

where $\mathcal{A}_c$ represents the pre-softmax attention value, $\mathcal{A}_{score}$ represents the attention score, d represents the hidden size of the attention head, and $\mathcal{A}$ presents the attention output of the head. The proposed novel multi-head self-attention with L1 norms-based residual connection is illustrated in Fig. 2 and is explained in the following steps.

1) The scaled Dot-Product Attention Block (Fig. 2(a)) is made to produce two outputs: attention score, $\mathcal{A}_{score}$ which is the softmax of the scaled value of matrix multiplication of Q & K^T and attention output, $\mathcal{A}$ which is the matrix multiplication of attention score and the value (V) as given by Eqs. (2) and (3), respectively. Both the outputs from all such attention heads are collected.
2) The attention outputs are concatenated and passed through a linear projection layer as in the original version of ViT.
3) L_1 Norms of the attention value ($\mathcal{A}_c = QK^T/\sqrt{d}$) from all attention heads are computed and sorted. The attention score with the highest norm value is treated as the attention head having the highest attention probability. The attention output ($MatMul$ of the attention score and the value (V) terms) corresponding to the best attention probability is collected. This attention output is added as the residual term to the final projected attention output by the linear projection layer, shown in Fig. 2(b). If the number of heads in the multi-head attention block is H, then the final attention output is computed by using Eqs. (4) and (5)

$$j = \arg\max_{i=1:H} \|\mathcal{A}_c(i)\|_1 \tag{4}$$

$$\mathcal{A}_{final} = LinearProj\left(\left[\mathcal{A}_1\ \mathcal{A}_2\ \cdots \mathcal{A}_H\right]\right) + \mathcal{A}_j \tag{5}$$

where $LinearProj(.)$ represents the linear projection of the concatenated output to convert the dimensionality back to the same dimensionality as the output of a single head.

The inclusion of a sorting mechanism, even though a non-differentiable step, does not produce gradient error during the backpropagation for model weight updates. PyTorch-style autograd works based on the computational graph traced during the forward pass. Once the best attention head is selected in the forward pass, the computational graph is fixed, and only the selected tensor is part of the graph. Therefore, backpropagation of the residual connection flows only through the selected head, and Other heads (non-selected) receive no gradient via residual.

MLP block: MLP block is a non-linear feedforward neural network that processes the token embedding and extracts higher-level features from the attention output. We have used two linear layers with a non-linear activation function in between them to construct the MLP block. The first linear layer applies a linear transformation to the input features, mapping them to a higher-dimensional space of 4 times the embedding dimension. Next, a non-linear activation function is applied element-wise to the transformed features. This introduces non-linearity into the model and allows it to capture more complex relationships between the input features. The output of the non-linear activation layer is given to the second linear layer, which maps the features back to the original embedding dimension.

4. Experimental results and discussion

The efficacy of the proposed framework of ViT with selective residual in multihead self-attention is demonstrated on bearing fault diagnosis datasets: (i) CWRU fault diagnosis bearing data [30] and (ii) Paderborn University dataset [31].

4.1. Dataset description

4.1.1. *CWRU Bearing data:*

The bearing dataset provided by CWRU [30] was recorded on a ball bearing testing platform. Using electro-discharge machining, the motor bearings were seeded with fault diameters of 7, 14, and 21 mils (1 mil = 0.001 in.). For each case of the fault diameter, defects were created at the inner raceway, rolling element (i.e., ball), and outer raceway of the motor bearing. The vibration signal represents four different states of the machine: (a) healthy (Normal: **N**), (b) inner race (**IR**), (c) outer race (**OR**) and (d) rolling element (ball: **B**). The vibration data were recorded for each fault diameter with motor loads of 0 to 3 hp and motor speeds of 1730 to 1797 RPM. The datasets were recorded under four different cases of fault location and sampling frequency:

a) Normal baseline data recorded at $12k$ samples/sec
b) Drive end (DE) fault data with $12k$ samples/sec
c) Drive end (DE) fault data with $48k$ samples/sec
d) Fan-End (FE) bearing fault data (recorded at $12k$ samples/second)

4.1.2. *Paderborn university (PBU) dataset:*

The Paderborn University dataset [31] is the best dataset for the analysis of bearing faults on electromagnetic rotating machines under a wide variety of operating conditions. The vibration signals were recorded by performing 32 different bearing experiments categorized as

i) 6 different experiments were conducted on healthy bearings.
ii) 12 different experiments were conducted on artificially damaged bearings.
iii) 14 Different experiments were conducted on real damaged bearings by accelerated lifetime tests.

Datasets from each experiment have measurements of motor phase currents, vibration, speed, torque, bearing temperature, and radial force. Each dataset contains 20 measurements of 4 seconds under four different settings of speed, torque, and force, termed as load settings. These four load settings are (i) **L1: N09_M07_F10** (speed = 900 rpm, torque = 0.7 Nm & radial force = 1000 N), (ii) **L2: N15_M01_F10** (speed = 1500 rpm, torque = 0.1 Nm & radial force = 1000 N) (iii) **L3: N15_M07_F04** (speed = 1500 rpm, torque = 0.7 Nm & radial force = 400 N) and (iv) **L4: N15_M07_F10** (speed = 1500 rpm, torque = 0.7 Nm & radial force = 1000 N). There are two classes of faults: **inner race (IR) damage** and **outer race (OR) damage**. Each class of fault encompasses a wide variety of damages, with different levels of damage represented by the extent of the damage. The details of different bearing names, fault level (extent), fault class, and different settings of speed, torque, and force can be found in [31].

4.2. Dataset pre-processing

The fault diagnosis data contains time-series samples recorded for a fixed time interval. For the training of the proposed framework, the recorded time-series signals are converted into graphical representations of frequency transforms, called short-time Fourier transforms (STFT). Fast Fourier transforms (FFT) are applied sequentially on the windowed time-series data to obtain time-localized frequency patterns [32]. For this conversion, the Python library of *matplotlib.axes.Axes.specgram* has been applied with a fixed window size of 100 samples for CWRU and 400 for the PBU dataset, the number of data points used in each block for the FFT (NFFT) of 32, and the sampling frequency (Fs) of 1000. The spectrogram representation of n samples (segmented) with 100 data points has been shown in Fig. 3.

4.3. Evaluation scheme

1. **Source Data for pre-training** The source dataset is prepared from *12k Hz Drive End (DE)* fault with *fault diameter* = 7 *mil and load* = 0 *hp*. The recorded sample of length 121000 from each class is used. Therefore, the time series data with data points 121000 is converted into 1210 images per class by applying STFT with a window length of 100 points. The combined dataset, including all four classes (N:IR:B:OR) contains a total of 4840 sample images.
2. **Target-1:** The target-1 (T1) is prepared using the time-series signals recorded at Fan End (FE) for the *12k Hz* with 21 mil fault diameters and at 1, 2, & 3 hp motor loads. Therefore, a total of three different sub-cases (datasets) are created using different conditions of the motor load, each having 4 classes: N:IR:B:OR. For each case, $40,000$ data points per class from the time-series signal are used to create 400 images per class, assuming 100 points as window size for STFT conversion.

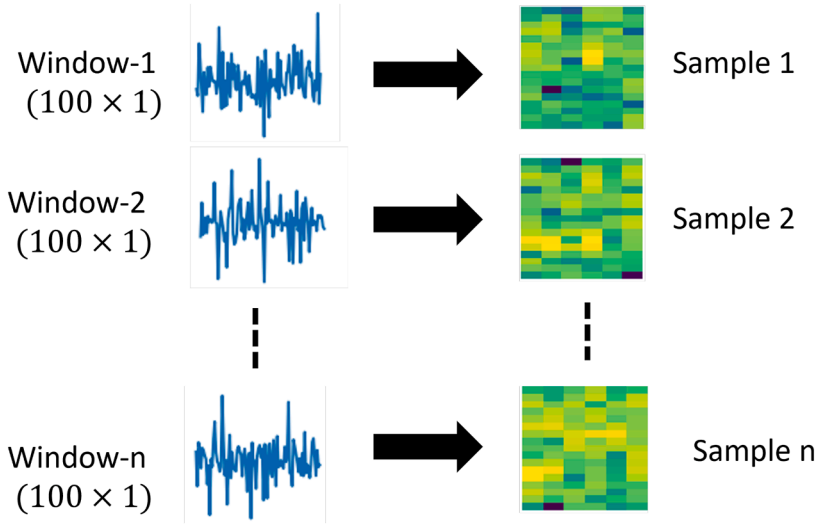

Fig. 3. Sample examples of STFT-based conversion of time-series signal into spectrogram images with window size = 100.

3. **Target-2:** Another group of target datasets, target-2 (T2), is created under four different operating conditions of load and torque settings named L1, L2, L3, & L4. For each case, 200,000 data points from one measurement file are used to create 500 images per class with 400 data points as window size for STFT conversion.
4. **Target datasets with AWGN** The time series data samples from both the target datasets T1 and T2 are now corrupted with AWGN having strength such that the resulting signal has SNR = 10dB. Then, STFT conversion is applied with the same time window to obtain the same number of images per class as mentioned above for the cases T1 and T2.

For the above target cases, the dataset was split into train, test, and validation using a random sampling method.

4.4. Comparison with state-of-the-art methods

The proposed framework of ViT with the selective residual in the multi-head self-attention block is compared with state-of-the-art methods for image classification. The state-of-the-art methods selected are listed as Support Vector Machine (SVM) as baseline method [33], SqueezeNet [34], GoogleNet [35], DenseNet [36], ResNeXt [37], EfficientNet [38], Vanilla ViT [18] (base model with 12 encoder blocks: "B_16_imagenet1k", and ReViT [24].

The SVM model is selected as a baseline model for comparison and is trained using flattened features obtained using the spectrogram images to maintain consistency in the image-based diagnosis. For all other models, pre-trained weights are downloaded from the PyTorch hub, customized to make them compatible with the number of classes in the target dataset, and then fine-tuned on the target dataset cases: T1, T2, and with AWGN with SNR = 10dB.

4.5. Evaluation metrics

The performance of a diagnostic model is measured in terms of the following evaluation metrics:

1. **Classification accuracy (CA)**: Classification accuracy (CA) as widely accepted in the literature [7,39,40]. CA is defined as

$$CA = \frac{\text{Number of correct classifications}}{\text{Total number of test samples}} \times 100\,\% \tag{6}$$

2. Transfer Improvement (TI): TI defines the relative improvement of the performance of a model with respect to a baseline method. TI in terms of average CA is calculated as $TI = \overline{CA} - \overline{CA}_{baseline}$, where, $\overline{CA}$ is the average CA for datasets under various operating conditions.
3. Confusion Matrix: The confusion matrix is the graphical representation of classification performance in terms of the number of correct and missed classifications arranged as $c \times c$ matrix, where c represents the number of classes.

4.6. Implementation details

The proposed ViT with the novel multi-head self-attention mechanism, as shown in Fig. 2 is implemented using the PyTorch library of Python on the Google Colaboratory platform. The multi-head self-attention with residual connection, as depicted in Fig. 2 is implemented as a Python class definition using the nn-module of PyTorch. All other blocks are defined similarly to the standard ViT model. The classifier head is defined for four classes present in the source dataset: N:IR:B:OR. The model's hyperparameters were selected as follows: number of encoder blocks = 12, number of heads in multi-head attention blocks = 4, patch size = 16×16, input image size = 128, and number of hidden sizes for MLP = 768. The weights of the proposed model are initialized using Xavier initialization. Then, the model is trained with source data described in 4.3 for 40 epochs on the NVIDIA A100-SXM GPU from the Google Co-Laboratory.

The model trained on the source dataset serves as a pre-trained model for fine-tuning models for other cases of the target dataset. Therefore, the weights of the model trained on the source dataset are saved as a weight dictionary. Now, the models for the target datasets are trained using the following steps:

i) Import all the definitions required for the proposed model's implementation.
ii) Load the pre-trained (source) model from the saved dictionary.
iii) Re-define the classifier head to make it suitable for the number of classes in the target dataset. For target datasets in T1, the number of classes remains the same, which is equal to 4. For target datasets in T2, the classifier head is redefined for three classes.
iv) Now the model with the new classifier head is fine-tuned with the target samples for 40 epochs for both cases of target datasets: (i) T1, & T2, and (ii) T1, & T2 with AWGN (SNR = 10dB). The training statistics are evaluated in terms of $\%CA$ with respect to epochs on the training and validation datasets. The comparisons of the validation $\%CA$ curves for one of the PBU datasets (T2-L1) are shown in Fig. 4.

4.7. Results

The performance in terms of CA for the proposed ViT model and the aforementioned state-of-the-art models on target datasets T1 and T2 with and without noise is presented in Tables 1 and 2. Table 1 contains the performance comparisons of all the aforementioned models on the three different operating conditions of the CWRU dataset (T1) for both cleaned and noisy datasets. Similarly, Table 2 contains the $\%CA$s of all models for four different cases from the PBU dataset under cleaned and noisy cases.

Apart from the performance evaluation in terms of $\%CA$, the classification performance has also been evaluated using the confusion matrix. The confusion matrices for all the models on one of the target datasets (T2-L1) and the same data with added noise have been shown in Figs. 5 and 6. It can be observed that the proposed framework performed well with even the noisy dataset, with minor misclassification. Fig. 7 compares the TI charts in terms of $\overline{CA}$, calculated overall sub-cases of (a) CWRU Dataset and (b) CWRU Dataset with AWGN (SNR = 10dB). Similarly, Fig. 8 shows the TI charts in terms of $\overline{CA}$ for (a) PBU Dataset and (b) PBU Dataset with AWGN (SNR = 10 dB).

4.8. Ablation study

An ablation study was conducted to analyze the effect of the following aspects of the proposed framework: (i) the norm to select the best attention score for residual in multihead attention, (ii) the window size for STFT, and (iii) the number of data points in each block of the FFT

4.8.1. Effect of norm

L_1 Norm is applied in the proposed framework to select the best attention head to encourage the head specialization under sparse distributions of multihead learning. To further strengthen the generalization, the proposed framework was trained on all cases of the target sets (with and without noise) by replacing L_1-norm with L_2-norm (Euclidean norm) and L_∞-norm. The classification performances are shown in Table 3.

It can clearly be observed that the model works better with the L_1-norm, also known as the Manhattan Norm, in the current scenario of attention scores of input patches. The L_1 norm is applied here in grid-based systems, which helps the multihead attention module to capture the best grid patterns produced by the individual head. In the case of other norms, the attention scores of heads are flattened along the spatial dimensions to apply the norm. The spatial geometric grid is completely lost; therefore, the model does not benefit from head specialization due to the skip connection of the best attention.

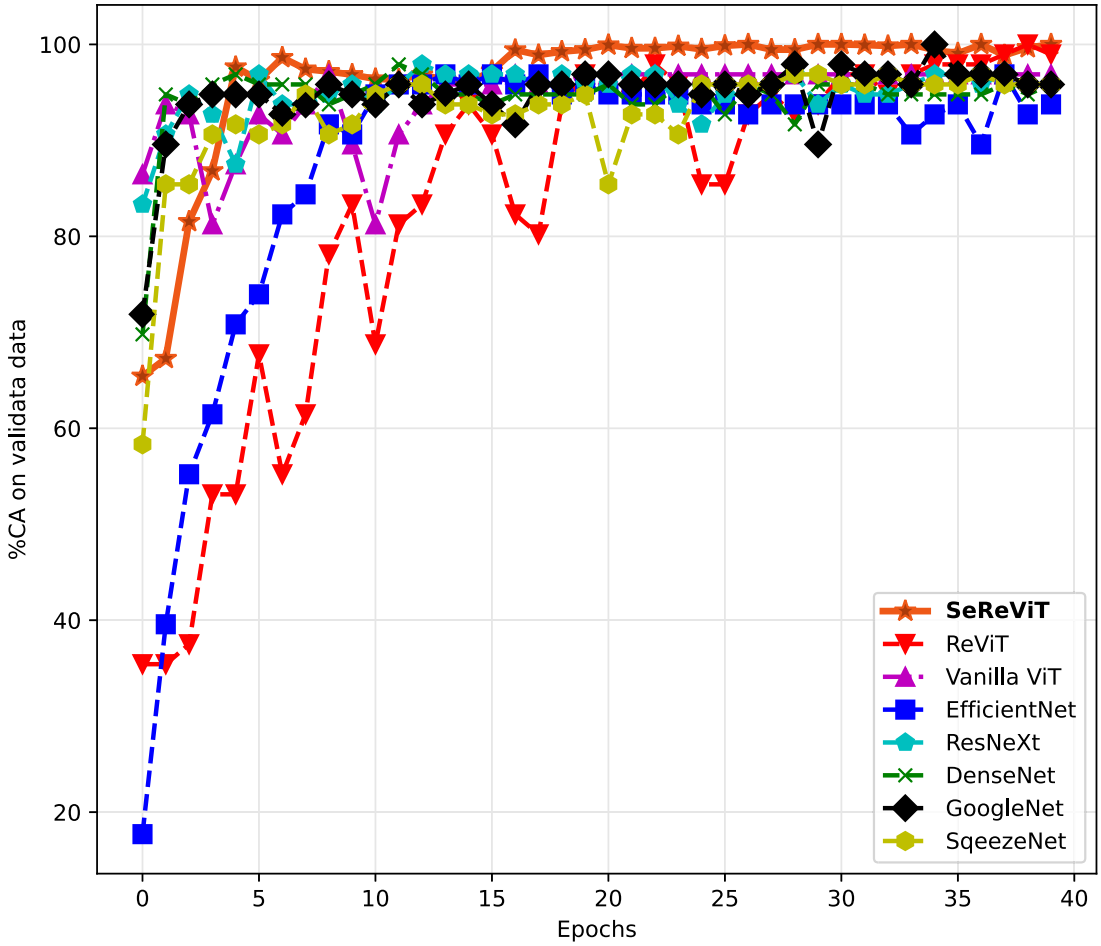

Fig. 4. Comparison of validation $\%CA$ curves for the dataset case: T3-L1.

Table 1
Diagnostic performance in term of CA on the datasets T1.

Target	Load	SVM [33]	SqueezeNet [34]	GoogleNet [35]	DenseNet [36]	ResNext [37]	EfficientNet [38]	Vanilla ViT[18]	ReViT [24]	**SeReViT (Proposed)**
T1	1 hp	95.9	94.4	95.6	95.3	95.3	94.7	97.5	99.1	**99.7**
	2 hp	86.3	91.3	90.6	93.8	91.3	96.3	98.1	97.5	**99.7**
	3 hp	87.5	87.5	88.1	87.5	90.3	90.9	96.3	98.8	**99.4**
T1 (SNR = 10dB)	1 hp	88.8	88.4	87.8	90.6	89.4	90.3	94.7	96.3	**98.4**
	2 hp	81.3	86.3	90.6	86.3	90.6	91.3	91.3	97.5	**98.8**
	3 hp	78.1	81.3	81.3	86.3	83.1	86.3	92.4	97.8	**98.1**

4.8.2. *Effect of window size and data points in STFT conversion*

The window size is selected based on the segment length selected for the preparation of training samples in [41]. The segment length provides the basis to create the samples from the time-series data with features equal to the length of the segment. The size of the segment is best decided based on the number of data points recorded in 1/4th of the time taken for one rotation of the bearing shaft [30,31]. Therefore, the window sizes for CWRU data and the Paderborn University data are selected as 100 and 400, respectively [41]. However, we validated our model by varying the window size and the number of data points (NFFT) to further generalize the proposed framework. Experimentations by varying the window size and the value of NFFT were performed on the first case from the target-1 dataset (T1-1hp). The classification performance is shown in Table 4.

This experimentation clarifies that changing the number of data points in each block of NFFT has a very negligible effect, provided that the number of overlaps is less than the size of NFFT, a required condition for STFT. The variations in the window size change the number of samples created. Additionally, if data points recorded in 1/4th of the time taken for one rotation are used as one sample, it better represents the fault signature [41]. Therefore, the window size of 100 for CWRU performs better.

4.9. *Discussion*

The classification performances of the proposed framework and selected state-of-the-art methods lead to the following observations:

i) The performance comparisons in Tables 1 and 2 for target datasets T1 and T2 with and without AWGN (SNR = 10dB) reveal that the accuracy and the reliability of image-based diagnosis have been significantly improved by adding the residual of the best attention output as described in Section 3. The accumulative attention enables the model to capture contextual information even with the spectrogram images obtained with noisy data.

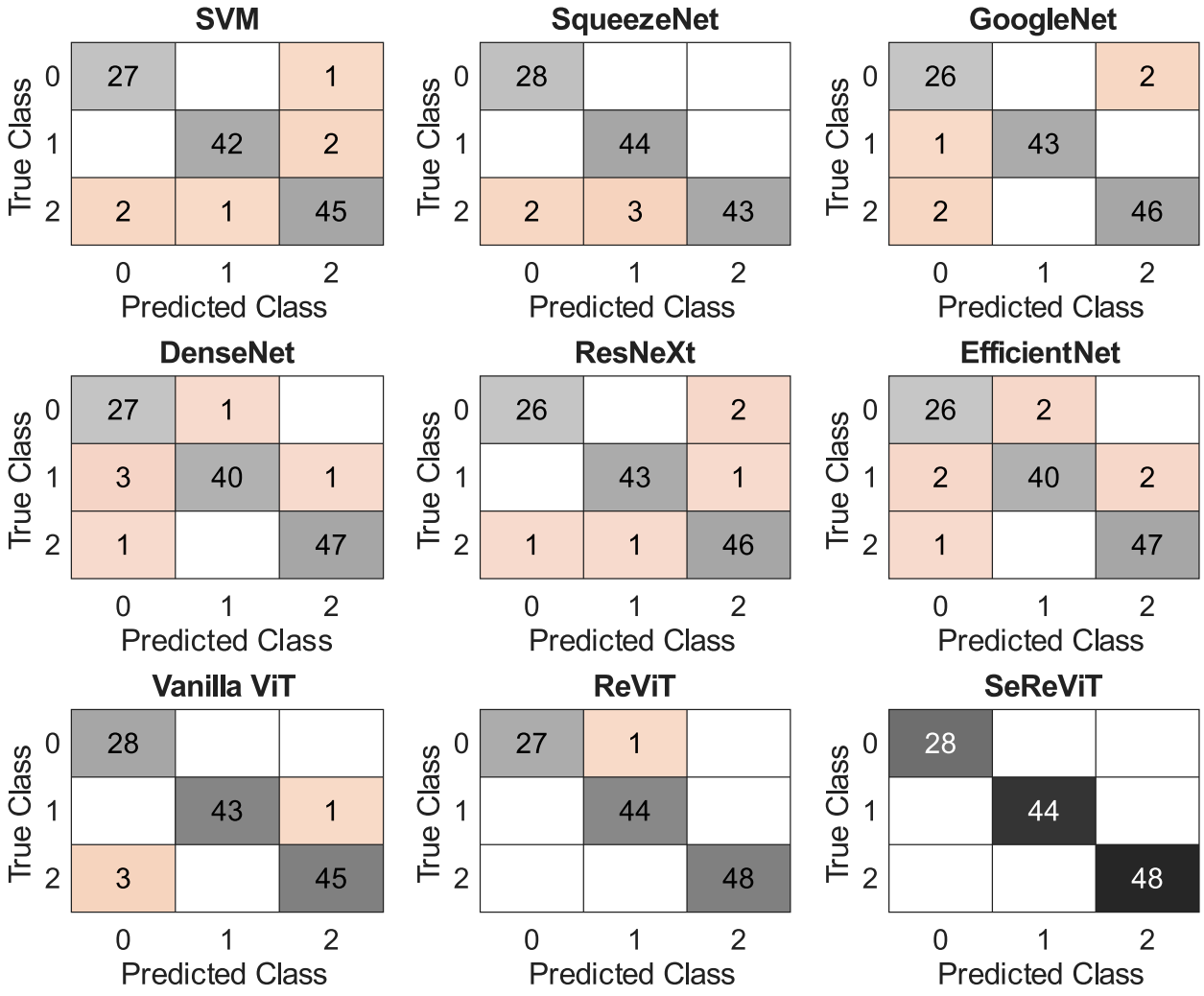

Fig. 5. Confusion matrices for the dataset case T4-L1; class label {'0', '1', '2'} represents the class name {'H', 'OR', 'IR'}.

Table 2
Diagnostic performance in term of CA on the target datasets from T2.

Target	L.S.	SVM [33]	SqueezeNet [34]	GoogleNet [35]	DenseNet [36]	ResNext [37]	EfficientNet [38]	Vanilla ViT[18]	ReViT [24]	**SeReViT (Proposed)**
T2	L1	94.7	95.83	95.83	95.3	95.83	93.75	96.88	98.96	**100**
	L2	93.3	94.7	91.7	95	95.7	97.3	96.3	98.3	**99.7**
	L3	93.3	93.3	90.7	94.3	96.7	96.7	94.3	99.7	**100**
	L4	90.7	90.7	94.7	91.7	93.3	92.3	96.7	98.7	**99.3**
T2 (SNR = 10dB)	L1	95	90	89.7	93	90.7	93.7	96.3	97.7	**98.3**
	L2	78.7	93.3	87.5	91.7	90	95.3	93.3	95.3	**97.7**
	L3	93.3	87.5	90.3	90.7	93.7	95	91.7	98	**98.3**
	L4	87.5	91.3	89.3	87.5	91.3	91.7	93.3	96.7	**97.3**

Table 3
Effect of various norms on the performance (%CA) on the target datasets from group T1 and T2.

Method	Target Datasets													
	T1			T1 (SNR = 10 dB)			T2				T2 (SNR = 10 dB)			
	1hp	2hp	3hp	1hp	2hp	3hp	L1	L2	L3	L4	L1	L2	L3	L4
SeReViT with L_1-Norm	99.7	99.7	99.4	98.4	98.8	98.1	100	99.7	100	99.3	98.3	97.7	98.3	97.3
SeReViT with L_2-Norm	98.8	97.5	97.5	98.1	97.5	96.3	97.7	94.3	94.3	96.7	91.7	92.3	90	91.7
SeReViT with $L\infty$-Norm	97.5	96.3	94.7	97.5	94.7	96.3	96.7	96.9	98.3	97.7	95.3	95.3	97.7	95.3

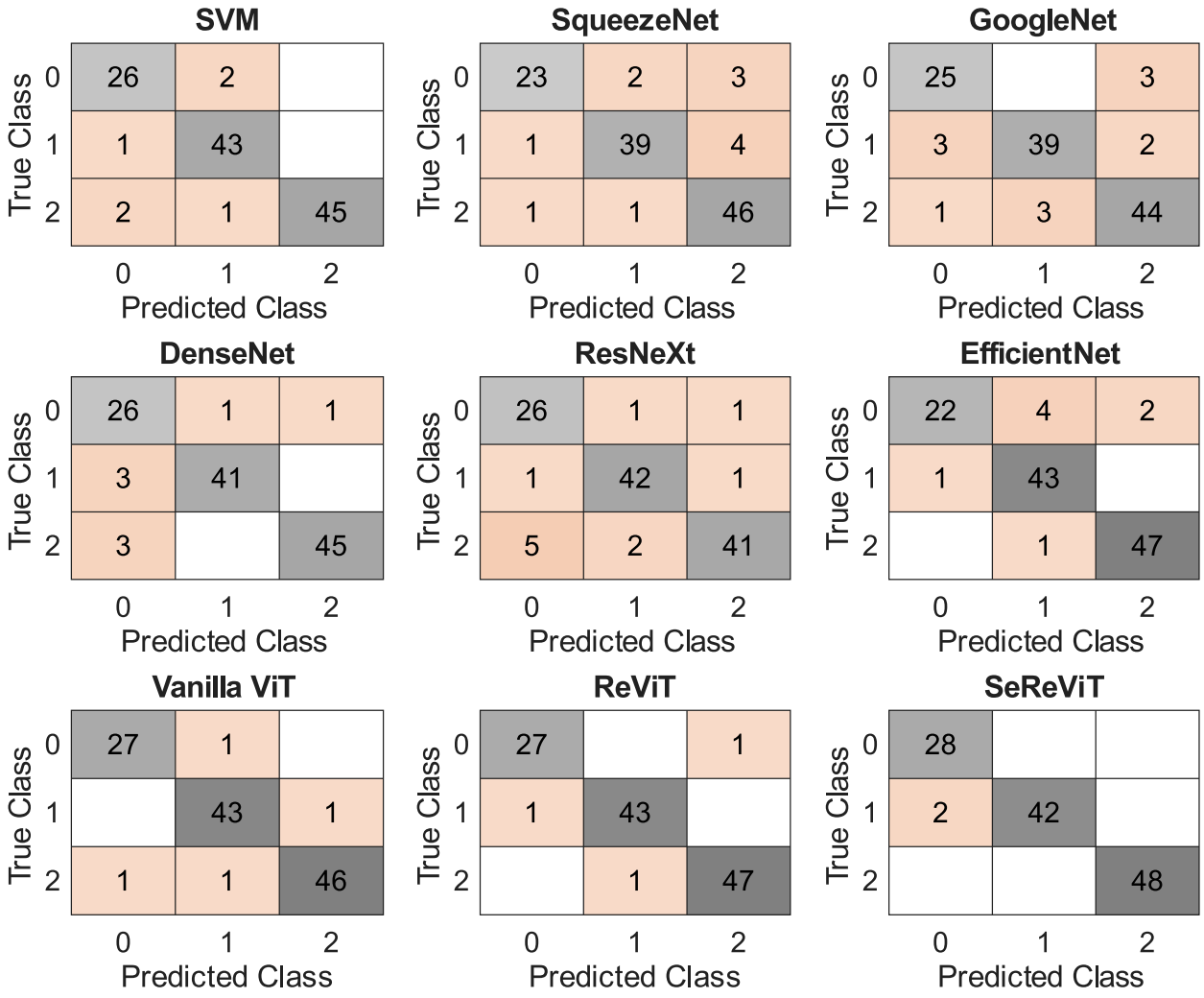

Fig. 6. Confusion matrices for the dataset case T4-L1 with AWGN (SNR = 10dB); class label {'0', '1', '2'} represents the class name {'H', 'OR', 'IR'}.

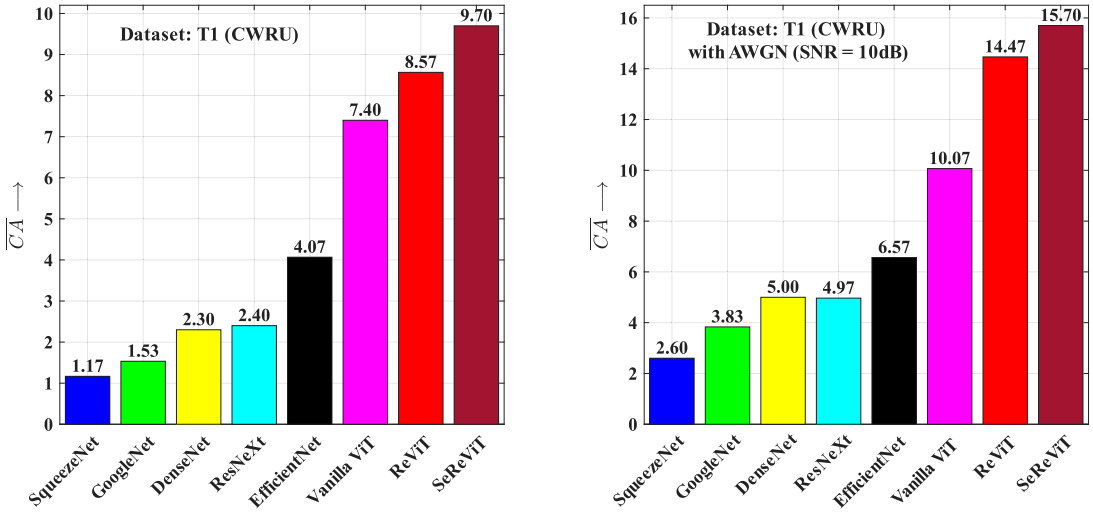

Fig. 7. TI charts in terms of $\overline{CA}$ for (a) CWRU Dataset and (b) CWRU Dataset with AWGN (SNR = 10dB).

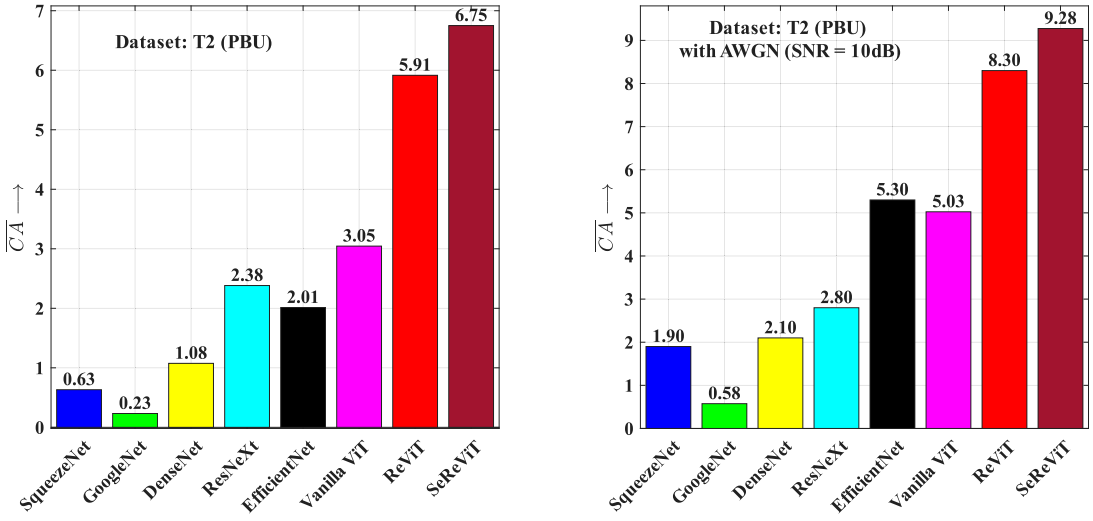

Fig. 8. TI charts in terms of $\overline{CA}$ for (a) PBU Dataset and (b) PBU Dataset with AWGN (SNR = 10dB).

Table 4
Effect of window size and NFFT on the performance (%CA) on the target dataset T1, 1hp.

Window size	50			100			200		
NFFT	16	32	64	16	32	64	16	32	64
Samples/class	800	800	800	400	400	400	200	200	200
%CA	95.6	96.3	94.7	**99.7**	**99.7**	99.4	90.3	91.3	91.3
%CA with Noise	91.3	92.4	91.3	**98.4**	**98.4**	98.1	86.3	86.3	87.8

ii) The training performance of most of the models varies from dataset to dataset. The validation curves for the various models for one of the target data cases (T2-L1) show the faster convergence and stability of the proposed model compared to all other models.

iii) The confusion matrices shown in Figs. 5 and 6 for the target dataset case T2l-L1, with and without noise, respectively, demonstrate the classification report against the ground truth. The numbers in the diagonal boxes provide the number of correct classifications, whereas the off-diagonal numbers show misclassifications. It can be observed that there are very few misclassifications by the proposed framework, even with added noise. Therefore, the proposed framework is more robust against the input noise in the recorded vibration signal.

iv) TI charts shown in Figs. 7 and 8 illustrate the overall improvement of the proposed framework compared to all the state-of-the-art models for image classification, assuming SVM as the baseline method. It can be noted that the performance of some of the models is even poorer than the baseline method. However, in the case of datasets under noisy conditions, all the deep learning models have better performance, which shows that SVM is much more affected under noisy conditions.

v) Overall, it can be concluded that the proposed framework of ReViT and the training mechanism provide sufficient model learning even under noisy conditions. The fine-tuned model with the target dataset cases performs up to 100 % classification accuracy; however, the computational complexity in terms of feed-forward response time for testing would be much higher than that of fully connected models obtained by using EvoN2N [10] and GS-EvoN2N [11] for a similar application of fault diagnosis. Therefore, the proposed framework can be suitable for image-based diagnosis in applications to avoid the need for signal processing/filtering for noise removal.

4.10. Complexity analysis

The complexity of the multi-head attention mechanism in traditional vision transformers per layer is $O(n^2 \cdot d)$, where n and d represent the sequence length and dimension of the heads, respectively [18]. In our proposed framework, we introduce a sorting step for the L-1 norms of attention scores in each layer. Computing the L-1 norms for a matrix of size $n \times n \times d$ has a complexity of $O(n \cdot n \cdot d)$, and the complexity for sorting the L-1 norms having a size equal to the number of heads d is $O(d \cdot \log(d))$. Therefore, the total complexity of ViT with the proposed attention mechanism becomes $O(n^2 \cdot d) + O(n^2 \cdot d + d \cdot \log(d))$. By ignoring the summation term with lower complexity, the overall time complexity of the proposed framework is given by $O\left(n^2 \cdot d + d \cdot \log(d)\right)$.

5. Conclusions

In this research work, we propose a novel framework of selective residual in multihead self-attention for the Vision Transformer (ViT). The proposed framework utilizes the Manhattan (L-1) norms of attention scores to identify the best attention output from the multiheads. The attention scores are sorted according to L-1 norms to get the highest attention score and the corresponding attention head index accordingly. The attention output of that particular head is carried forward to add to the final projected attention as a residual. Thus, the selective residual provides the model to focus the training on the specialized attention head. The validation results on the fault diagnosis datasets taken from the CWRU dataset and the PBU dataset under variable operating conditions, as well as with and without noise, justify the robustness against the noise contamination. Also, it demonstrates that the proposed framework achieves superior diagnostic performance compared to the vanilla ViT model as well as ReViT. The performance on the noisy data with AWGN (SNR = 10 dB) shows that the proposed framework may be a promising solution for fault diagnosis using raw data, alleviating the need for noise filtering.

The inclusion of L_1 norm-based sorting in the feedforward just provides selective residual and does not affect the gradient propagation for the weight update; therefore, the mechanism can be applied for any other form of multihead attention mechanism. This work does not demonstrate and validate the application of similar residual connections

for other variants of the ViT models. This work can be extended to apply the sorting-based selective residual in other variants of ViT for improved performance and robustness. Also, in the current research work, we considered improving the capability of capturing contextual information in sequential dependencies. This work can further be extended to include mechanisms that can also capture spatial patterns and apply the selective residual.

CRediT authorship contribution statement

Arun Kumar Sharma: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Nishchal K. Verma:** Supervision.

Author contribution

Arun Kumar Sharma: Conceived the primary idea, implemented the methodology, and produced the results. critically examined the results, and wrote the manuscript.
Nishchal K. Verma: Supervision and reviewing the manuscript.

Data availability

Data are openly available and cited at suitable place

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary material

Supplementary material associated with this article can be found in the online version at 10.1016/j.patcog.2025.112497

References

[1] Y. Li, L. Zhang, P. Liang, X. Wang, B. Wang, L. Xu, Semi-supervised meta-path space extended graph convolution network for intelligent fault diagnosis of rotating machinery under time-varying speeds, Reliability Engineering & System Safety 251 (2024) 110363. https://doi.org/10.1016/j.ress.2024.110363

[2] J. Cen, W. Si, X. Liu, B. Zhao, C. Xu, S. Liu, Y. Xin, Diffusion model and vision transformer for intelligent fault diagnosis under small samples, Meas. Sci. Technol. 35 (3) (2023) 036204. https://doi.org/10.1088/1361-6501/ad179c

[3] Q. He, Y. Liu, K. Fanrang, Machine fault signature analysis by midpoint-based empirical mode decomposition, Meas. Sci. Technol. 22 (2010) 015702. https://doi.org/10.1088/0957-0233/22/1/015702

[4] N.K. Verma, R.K. Sevakula, S. Dixit, A. Salour, Intelligent condition based monitoring using acoustic signals for air compressors, IEEE Trans. Reliab. 65 (1) (2016) 291–309. https://doi.org/10.1109/TR.2015.2459684

[5] W. Fan, Q. Zhou, J. Li, Z. Zhu, A wavelet-Based statistical approach for monitoring and diagnosis of compound faults with application to rolling bearings, IEEE Trans. Autom. Sci. Eng. 15 (4) (2018) 1563–1572. https://doi.org/10.1109/TASE.2017.2720177

[6] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, V. Lempitsky, Domain-Adversarial training of neural networks, Journal of Machine Learning Research 17 (59) (2016) 1–35.

[7] W. Lu, B. Liang, Y. Cheng, D. Meng, J. Yang, T. Zhang, Deep model based domain adaptation for fault diagnosis, IEEE Trans. Ind. Electron. 64 (3) (2017) 2296–2305. https://doi.org/10.1109/TIE.2016.2627020

[8] X. Li, W. Zhang, Q. Ding, Cross-Domain fault diagnosis of rolling element bearings using deep generative neural networks, IEEE Trans. Ind. Electron. 66 (7) (2019) 5525–5534. https://doi.org/10.1109/TIE.2018.2868023

[9] L. Guo, Y. Lei, S. Xing, T. Yan, N. Li, Deep convolutional transfer learning network: a new method for intelligent fault diagnosis of machines with unlabeled data, IEEE Trans. Ind. Electron. 66 (9) (2019) 7316–7325. https://doi.org/10.1109/TIE.2018.2877090

[10] A.K. Sharma, N.K. Verma, Knowledge Transfer based Evolutionary Deep Neural Network for Intelligent Fault Diagnosis, arXiv:2109.13479 [eess.SP] (2021).

[11] A.K. Sharma, N.K. Verma, Guided sampling-based evolutionary deep neural network for intelligent fault diagnosis, Eng Appl Artif Intell 128 (2024) 107498. https://doi.org/10.1016/j.engappai.2023.107498

[12] Y. Xue, D. Dou, J. Yang, Multi-fault diagnosis of rotating machinery based on deep convolution neural network and support vector machine, Measurement 156 (2020) 107571. https://doi.org/10.1016/j.measurement.2020.107571

[13] P. Liang, C. Deng, J. Wu, Z. Yang, Intelligent fault diagnosis of rotating machinery via wavelet transform, generative adversarial nets and convolutional neural network, Measurement 159 (2020) 107768. https://doi.org/10.1016/j.measurement.2020.107768

[14] Y. Cheng, M. Lin, J. Wu, H. Zhu, X. Shao, Intelligent fault diagnosis of rotating machinery based on continuous wavelet transform-local binary convolutional neural network, Knowl Based Syst 216 (2021) 106796. https://doi.org/10.1016/j.knosys.2021.106796

[15] A. Choudhary, T. Mian, S. Fatima, Convolutional neural network based bearing fault diagnosis of rotating machine using thermal images, Measurement 176 (2021) 109196. https://doi.org/10.1016/j.measurement.2021.109196

[16] S. Liu, Z. Ji, Y. Wang, Z. Zhang, Z. Xu, C. Kan, K. Jin, Multi-feature fusion for fault diagnosis of rotating machinery based on convolutional neural network, Comput Commun 173 (2021) 160–169. https://doi.org/10.1016/j.comcom.2021.04.016

[17] T. Gonçalves, I. Rio-Torto, L.F. Teixeira, J.S. Cardoso, A survey on attention mechanisms for medical applications: are we moving toward better algorithms?, IEEE Access 10 (2022) 98909–98935. https://doi.org/10.1109/ACCESS.2022.3206449

[18] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020, 2010.11929

[19] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, 2021, 2012.12877

[20] Y. Barhoumi, G. Rasool, Scopeformer: n-CNN-ViT Hybrid Model for Intracranial Hemorrhage Classification, 2021, 2107.04575

[21] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, 2021, 2103.14030

[22] H. Wang, Y. Ji, K. Song, M. Sun, P. Lv, T. Zhang, ViT-P: Classification of genitourinary syndrome of menopause from OCT images based on vision transformer models, IEEE Trans Instrum Meas 70 (2021) 1–14. https://doi.org/10.1109/TIM.2021.3122121

[23] E.S. Sabry, S.S. Elagooz, F.E.A. El-Samie, W. El-Shafai, N.A. El-Bahnasawy, G.M. El-Banby, A.D. Algarni, N.F. Soliman, R.A. Ramadan, Image retrieval using convolutional autoencoder, infoGAN, and vision transformer unsupervised models, IEEE Access 11 (2023) 20445–20477. https://doi.org/10.1109/ACCESS.2023.3241858

[24] A. Diko, D. Avola, M. Cascio, L. Cinque, Revit: enhancing vision transformers feature diversity with attention residual connections, Pattern Recognit 156 (2024) 110853. https://doi.org/10.1016/j.patcog.2024.110853

[25] Y. Li, D. Chen, T. Tang, X. Shen, HTR-VT: Handwritten text recognition with vision transformer, Pattern Recognit 158 (2025) 110967. https://doi.org/10.1016/j.patcog.2024.110967

[26] X. Nie, H. Jin, Y. Yan, X. Chen, Z. Zhu, D. Qi, Scopevit: scale-Aware vision transformer, Pattern Recognit 153 (2024) 110470. https://doi.org/10.1016/j.patcog.2024.110470

[27] M. Rodrigo, C. Cuevas, N. García, Comprehensive comparison between vision transformers and convolutional neural networks for face recognition tasks, Sci Rep 14 (1) (2024) 21392.

[28] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, I. Titov, Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5797–5808. https://doi.org/10.18653/v1/P19-1580

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, 2017, 1706.03762

[30] W.A. Smith, R.B. Randall, ROLLING Element bearing diagnostics using the case western reserve university data: a benchmark study, Mech Syst Signal Process 64 (2015) 100–131.

[31] C. Lessmeier, J. Kuria Kimotho, D. Zimmer, W. Sextro, Condition Monitoring of Bearing Damage in Electromechanical Drive Systems by Using Motor Current Signals of Electric Motors: A Benchmark Data Set for Data-Driven Classification, in: Proceedings of the European Conference of the PHM Society 2016, European Conference of the Prognostics and Health Management Society, Bilbao (Spain), 2016. http://mb.uni-paderborn.de/kat/datacenter.

[32] N. Kehtarnavaz, CHAPTER 7 - Frequency Domain Processing, in: N. Kehtarnavaz (Ed.), Digital Signal Processing System Design (Second Edition), Academic Press, Burlington, 2nd edition, 2008, pp. 175–196. https://doi.org/10.1016/B978-0-12-374490-6.00007-6

[33] A. Widodo, B.-S. Yang, Support vector machine in machine condition monitoring and fault diagnosis, Mech Syst Signal Process 21 (6) (2007) 2560–2574. https://doi.org/10.1016/j.ymssp.2006.12.007

[34] F.N. Iandola, S. Han, M.W. Moskewicz, K. Ashraf, W.J. Dally, K. Keutzer, SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size, 2016, 1602.07360

[35] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going Deeper with Convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[36] G. Huang, Z. Liu, L. van der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, 2018, 1608.06993

[37] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, 2017, 1611.05431

[38] M. Tan, Q.V. Le, EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2020, 1905.11946

[39] M. Long, J. Wang, G. Ding, S.J. Pan, P.S. Yu, Adaptation regularization: a general framework for transfer learning, IEEE Trans Knowl Data Eng 26 (5) (2014) 1076–1089. https://doi.org/10.1109/TKDE.2013.111

[40] L. Wen, L. Gao, X. Li, A new deep transfer learning based on sparse auto-Encoder for fault diagnosis, IEEE Transactions on Systems, Man, and Cybernetics: Systems 49 (1) (2019) 136–144. https://doi.org/10.1109/TSMC.2017.2754287

[41] A.K. Sharma, N.K. Verma, Quick learning mechanism with cross-Domain adaptation for intelligent fault diagnosis, IEEE Transactions on Artificial Intelligence (2021) 1–1. https://doi.org/10.1109/TAI.2021.3123935