IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 November 2023, accepted 21 January 2024, date of publication 24 January 2024, date of current version 1 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3357946*

RESEARCH ARTICLE

# Exploiting Data-Efficient Image Transformer-Based Transfer Learning for Valvular Heart Diseases Detection

**TALIT JUMPHOO[1], KHOMDET PHAPATANABURI[2], WONGSATHON PATHONSUWAN[1], PATIKORN ANCHUEN[3], MONTHIPPA UTHANSAKUL[1], (Member, IEEE), AND PEERAPONG UTHANSAKUL[1], (Member, IEEE)**

[1]School of Telecommunication Engineering, Suranaree University of Technology, Nakhon Ratchasima 30000, Thailand
[2]Department of Telecommunication Engineering, Faculty of Engineering and Technology, Rajamangala University of Technology Isan (RMUTI), Nakhon Ratchasima 30000, Thailand
[3]Navaminda Kasatriyadhiraj Royal Air Force Academy, Muak Lek, Saraburi 18180, Thailand

Corresponding authors: Khomdet Phapatanaburi (khomdet.ph@rmuti.ac.th) and Peerapong Uthansakul (uthansakul@sut.ac.th)

This work was supported by in part by the Suranaree University of Technology (SUT); in part by the Thailand Science Research and Innovation (TSRI); in part by the National Science, Research and Innovation Fund (NSRF) through NRIIS under Grant 179284; and in part by NSRF via the Program Management Unit for Human Resources and Institutional Development, Research and Innovation (PMU-B), under Grant B13F660067.

**ABSTRACT** Recent studies have shown the potential of the Data-Efficient Image Transformer (DeiT)-based transfer learning method in speech/image recognition and classification utilizing models pre-trained on image datasets. However, the use of DeiT models, especially those pre-trained on image datasets, has not yet been explored for Valvular Heart Disease (VHD) detection. This paper proposes a transfer learning methodology using the DeiT model pre-trained on image datasets for VHD classification. Additionally, we introduce a hybrid Convolution-DeiT (Conv-DeiT) architecture to further improve classification performance. The Conv-DeiT framework integrates a convolutional block with a Squeeze-and-Excitation (SE) attention mechanism to enhance the channel and spatial information within the input features before processing by the DeiT model. The proposed models were assessed using the Heart Sound Murmur (HSM) database, accessible on GitHub. Experimental results show that the DeiT-based transfer learning approach achieved an overall accuracy of 97.44%. Moreover, our Conv-DeiT method outperformed the DeiT-based transfer learning with an impressive overall accuracy of 99.44%. This study indicates the effectiveness of transfer learning using DeiT models pre-trained on image datasets for heart sound classification. Specifically, our hybrid Conv-DeiT method, which combines the convolutional block and the SE-attention mechanism, demonstrates significant advantages in this context.

**INDEX TERMS** Valvular heart diseases detection, transfer learning, DeiT, hybrid model.

## I. INTRODUCTION

Valvular heart disease (VHD) is emerging as a major health concern globally, especially compared to other cardiovascular diseases, due to its rising prevalence and high mortality rates [1]. Early VHD screening is essential in reducing these mortality rates. Traditionally, auscultating heart sounds has been the primary medical approach for VHD evaluation, providing valuable insights into cardiovascular abnormalities [2], [3]. However, diagnosing cardiac abnormalities through auscultation can be challenging, especially for inexperienced clinicians [4]. In current technology, digital stethoscopes have been used to record heart sounds, which can be plotted in a graph known as Phonocardiograms

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

**TABLE 1. Summary of studies employing ML, DL, and transfer learning for heart sound classification: predominantly focused on five-class prediction using the Heart Sound Murmur (HSM) Database [62] and two-class prediction (Normal/abnormal) using the PhysioNet/CinC Database [56] and PASCAL Classifying Heart Sounds Challenge (PASCAL) [19].**

| Reference | Database | Features | Classification technique | Domain | Accuracy (%) |
|---|---|---|---|---|---|
| [5] | The HSM database | MFCC | KNN | ML | 84.30 |
| | | MFCC | MLP | ML | 80.90 |
| | | MFCC | SVM | ML | 86.10 |
| | | MFCC | DNN | DL | 89.10 |
| | | MFCCs based on EMD | KNN | ML | 88.40 |
| | | MFCCs based on EMD | MLP | ML | 88.80 |
| | | MFCCs based on EMD | SVM | ML | 96.20 |
| | | MFCCs based on EMD | DNN | DL | 98.90 |
| [6] | PhysioNet/CinC | Statistical, Frequency | XGBoost | ML | 92.90 |
| [7] | PhysioNet/CinC | Time, MFCCs and Statistical | Neural network (NN) | ML | 93.33 |
| [11] | PhysioNet/CinC | MFCC | Decision tree | ML | 86.40 |
| [12] | PhysioNet/CinC | MFCCs | NN | ML | 92.00 |
| [13] | PhysioNet/CinC | LPC, Entropy, MFCCs, DWT and PSD | NN | ML | 91.50 |
| [14] | Private | MFCC | DNN | DL | 91.12 |
| [15] | PhysioNet/CinC | MFCCs | DNN | DL | 93.00 |
| [16] | The HSM database | TQWT, EMD and Shannon energy | RBF neural networks | ML | 98.48 |
| [17] | PhysioNet/CinC | MFSCs | SVM | ML | 92.00 |
| [18] | PhysioNet/CinC | Gram polynomial and Fourier transform | NN | ML | 94.00 |
| [19] | PASCAL | DWT | Hidden Markov Models | ML | 92.74 |
| [20] | PhysioNet/CinC | Wavelet | CNN | DL | 81.20 |
| [21] | PhysioNet/CinC | Modified frequency slice wavelet transform | CNN | DL | 94.00 |
| [22] | PhysioNet/CinC | Frequency spectrum, Energy and Entropy | SVM | ML | 88.00 |
| [23] | Private | EMD and MFCCs | SVM | ML | 91.00 |
| [24] | MIT heart sounds | Frequency | SVM | ML | 98.00 |
| [25] | PhysioNet/CinC | Time, MFCC, DWT and Wavelet | SVM | ML | 82.40 |
| [26] | The HSM database | FMFE + MFCC | SVM | ML | 99.00 |
| [27] | PhysioNet/CinC | Spectral | SVM | ML | 98.00 |
| [28] | The HSM database | Time-frequency | MCC | ML | 98.33 |
| [29] | PhysioNet/CinC | Cochleagram | MLP | ML | 95.00 |
| [30] | The HSM database | Time-frequency magnitude and phase | RF | ML | 95.12 |
| [31] | Private | MFCC | KNN | ML | 98.00 |
| [32] | Private | EMD | KNN | ML | 94.00 |
| [33] | PASCAL | MFCCs | KNN | ML | 97.00 |
| [34] | Private | MFCCs | KNN | ML | 92.60 |
| [37] | PhysioNet/CinC | MFCCs | LSTM | DL | 98.61 |
| [38] | PhysioNet/CinC | Time, Frequency and Time–frequency | DNN | DL | 92.60 |
| [39] | PhysioNet/CinC | MFCC+ MFSC | 2D-CNN | DL | 81.50 |
| [40] | PhysioNet/CinC | Mean, Standard deviation and Power spectrum | CNN | DL | 86.02 |
| [41] | PhysioNet/CinC | Spectrogram, Mel-spectrogram and MFCCs | CNN | DL | 86.05 |
| [42] | The HSM database | Normalized signals | CNN | DL | 97.00 |
| [43] | PhysioNet/CinC | Spectrograms | 2D-CNN | DL | 97.05 |
| [44] | PhysioNet/CinC | - | 1D-CNN | DL | 93.28 |
| [45] | Private | - | 1D-CNN | DL | 93.56 |
| [46] | PhysioNet/CinC | Spectrograms | 1D-CNN | DL | 96.48 |
| [47] | The HSM database | Log-mel spectrogram | LSTM | DL | 95.00 |
| | | Log-mel spectrogram | CNN | DL | 99.67 |
| [48] | PhysioNet/CinC | LPCC and MFCC | SFF-HLSTM | DL | 99.10 |
| | | LPCC and MFCC | PFF-HLSTM | DL | 98.71 |
| [52] | PASCAL | MFCC | 2D-CNN + transfer learning | DL | 89.50 |
| [55] | The HSM database | FDPCT | VGG16-based transfer learning | DL | 94.00 |
| | | FDPCT | ResNet-50-based transfer learning | DL | 98.00 |
| | | FDPCT | Deep CNN | DL | 99.48 |

(PCG). It is paving the way for a comprehensive PCG signals database. As a result, using artificial intelligence to analyze and detect cardiac abnormalities has gained significant attention.

The current methods used in artificial intelligence for detecting VHD can be divided into two main methods: machine learning and deep learning. The machine learning approach is a process with manually designed feature extraction, converting the PCG signal into specific parameters, followed by a process that tunes learning features for a classifier to distinguish various VHD classes [5], [6], [7]. In contrast, the deep learning approach utilizes end-to-end systems that bypass manual feature extraction, leveraging deep learning-based classifiers to model and predict target classes [8], [9], [10].

Research on VHD detection mostly focuses on conventional pipeline approaches. These studies have explored the use of efficient hand-crafted feature extraction techniques in

combination with effective classifiers. Several methods have been proposed for the detection of VHD, including Mel-Frequency Cepstral Coefficients (MFCCs) [11], [12], [13], [14], [15], Tunable Q-factor Wavelet Transform (TQWT) [16], Mel Frequency Spectral Coefficients (MFSCs) [17], Gram polynomial [18] and Wavelet Transform (WT) [19], [20], [21]. In addition, the study examined various machine learning classifiers, including the support vector machine (SVM) [22], [23], [24], [25], [26], [27], multiclass composite classifiers (MCC) [28], Multi-layer Perceptron (MLP) [29], Random Forest (RF) [30], and k-Nearest Neighbor (k-NN) [31], [32], [33], [34]. The effectiveness of hand-crafted feature extraction for classification is a significant part of these conventional approaches, demonstrating the need for expertise, e.g., speech processing tasks [35].

According to previous research, deep learning models have demonstrated outstanding accuracy in detecting acoustic using complicated pattern recognition techniques [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46]. As reported by [5], the utilization of MFCC in Deep Neural Networks (DNNs) has proven to be more effective in detecting VHD than SVM, k-NN, and MLP classifiers. This superiority can be attributed to the DNN's inherent ability to independently extract hierarchical features from the MFCC. The utilization of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) models combined with a log-mel spectrogram has represented encouraging outcomes in detecting VHD [47]. Furthermore, a study by [48] proposed using hierarchical LSTM networks that incorporate parallel and series feature fusion to detect VHD through PCG signals. The study represents performance improvements when using MFCC and Linear Prediction Cepstral Coefficients (LPCC) as features. Nevertheless, the effectiveness of these deep learning classifiers is significantly influenced by the volume of training data [49].

Transfer learning-based classifiers have become more popular in many applications in recent times [50], [51], [52]. This technique leverages knowledge acquired from models trained on large datasets, which enhances data utilization efficiency and accelerates the training process [53]. According to [54], the approach enables models to achieve outstanding performance even when trained on a small amount of task-specific data, reducing the risk of overfitting. In addition, it enables the adaption of models from one domain to another one, resulting in enhanced performance in closely related domains. The studies operated by [55] focused on detecting VHD, which presented transfer learning models such as ResNet50 and VGGNet-16. These models were trained using Time-Frequency (TF) images derived from PCG signals, which results showed a promise in VHD classification. However, transfer learning may encounter difficulties when a substantial disparity exists between the source and target domains. This discrepancy necessitates the implementation of supplementary procedures to tackle these obstacles effectively. A detailed compilation of studies employing ML, deep learning, and transfer learning for the classification of heart sounds is summarized in Table 1, providing the reader with a more nuanced understanding of the field.

In this study, we explore the usefulness of the Data-Efficient Image Transformer (DeiT) model [57] in the context of heart sound classification using transfer learning. The inherent capabilities of DeiT include hierarchical feature extraction, and attention mechanisms, which help identify abnormalities in heart sounds. As a result, it is anticipated to be a promising candidate for detecting VHD. Although the architecture of DeiT models demonstrates exceptional performance in image domains, it is important to note that these models may not be fully optimized for the complexities associated with audio data. Therefore, it is imperative to enhance and modify the model to capture the intricacies of the heart more accurately.

In order to improve the classification efficacy of transfer learning based on DeiT, we propose integrating a hybrid Convolution-DeiT (Conv-DeiT) model. In order to enhance channel and spatial information before the VHD detection phase, we integrate a convolutional block and a squeeze-and-excitation (SE) attention mechanism into the DeiT framework. The present study provides the following contributions:

1) We explored the use of DeiT-based transfer learning for VHD detection. Specifically, we used three channels of the MFCC together with their corresponding delta and double delta coefficients, which are obtained from the original one-dimensional utterances. This approach enables us to use pre-trained DeiT models from image datasets.
2) We proposed a hybrid Conv-DeiT model to improve the classification efficacy of transfer learning based on DeiT. This model integrates a convolutional block and a SE-attention mechanism into the DeiT framework. It helps enhance the quality of channel and spatial input feature information before the DeiT and classification process.
3) In this study, we represented the capability of a DeiT model in the context of VHD detection, which was initially developed for image-based tasks. Moreover, the proposed Conv-DeiT technique enhanced VHD detection accuracy by integrating the convolutional block and SE-attention mechanism.

The subsequent sections of the paper are organized in the following manner: The proposed approach is described in Section II. The experimental setup and results are discussed in Sections III and IV, respectively. The last part provides commentary and suggests potential approaches for future research.

## II. PROPOSED METHOD

The proposed methodology for the detection of VHD relies on a framework of transfer learning. Initially, the audio signal undergoes a feature extraction phase where MFCCs and their derivatives are computed to capture both spectral and temporal dynamics. These features are then transformed

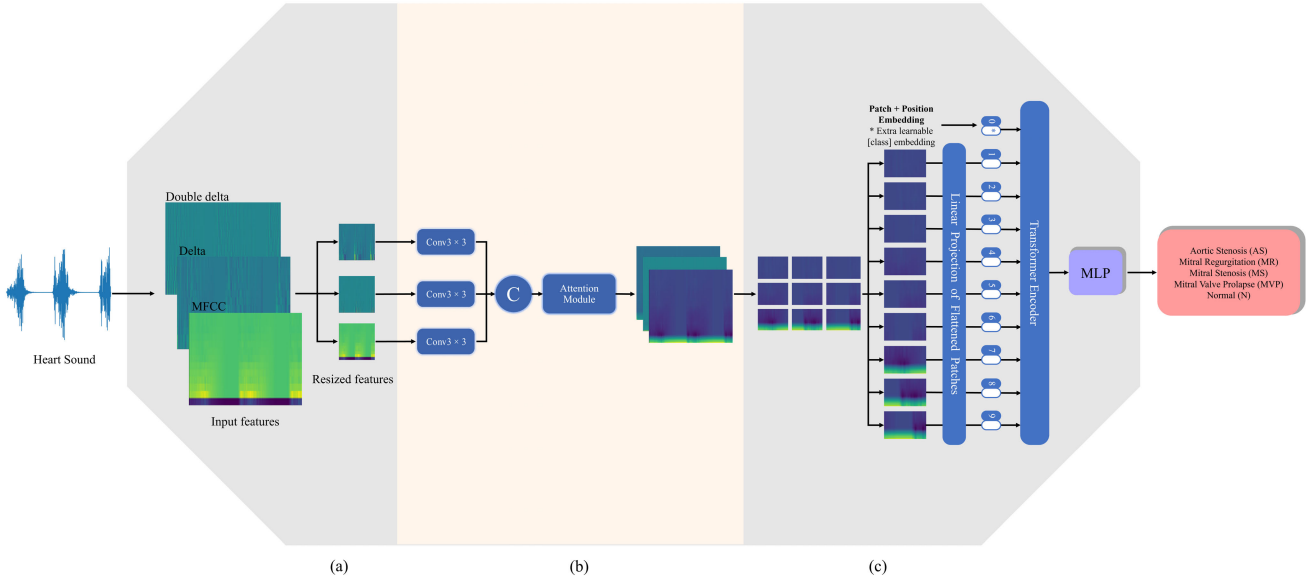



**FIGURE 1.** Overall visualization of our proposed architecture. (a) image-like feature extraction, (b) convolution module, and (c) classifier.

to conform to an image-like input suitable for our adapted Deep Image DeiT model, respecting its design for handling two-dimensional data structures. We have introduced a convolutional module consisting of parallel convolutional layers with self-attention mechanisms to augment the feature extraction process. This module is specifically engineered to adapt the spectral features of audio signals for the DeiT model, enabling it to leverage its spatial pattern recognition capabilities. The cohesive interplay and stepwise progression of these components are succinctly illustrated in Fig. 1. Furthermore, this section delves into the specific loss function employed in our approach. The following subsections provide a more granular breakdown of each component:

### A. TRANSFORMING AUDIO DATA FOR DeiT MODEL

The adaptation of the DeiT model from image to audio processing presents unique domain-specific challenges due to the fundamental differences between visual and auditory information representation. While DeiT excels in identifying patterns within the spatial domain of images, audio signals require an interpretation of patterns over both time and frequency domains. To address this, we transformed one-dimensional audio signals, denoted as $S$ into into a three-dimensional structure matching a $3 \times 224 \times 224$ image that parallel the two-dimensional nature of images, allowing us to leverage DeiT's powerful spatial pattern recognition capabilities. In order to achieve this, the MFCCs are employed.

$$MFCC = F(S) \tag{1}$$

MFCCs provide an audio representation that closely resembles the auditory perception of humans. The procedure entails dividing the audio into brief, overlapping segments, converting these segments into a frequency spectrum, and subsequently enhancing the frequencies that are most relevant to human auditory perception by employing the Mel scale. This concept is further enhanced by employing the Discrete Cosine Transform. Further details of the MFCC extraction can be seen in [58].

The temporal variations included in the MFCCs are effectively represented by incorporating the delta and double delta coefficients.

$$\Delta[t] = \sum_{n=1}^{N} n(MFCC[t+n] - MFCC[t-n]) \tag{2}$$

$$\Delta^2[t] = \sum_{n=1}^{N} n(\Delta[t+n] - \Delta[t-n]) \tag{3}$$

where $N$ represents the number of adjacent frames taken into account. Meanwhile, $n$ and $t$ stand for the frame's index and the current time frame.

The delta and double delta coefficients [59] adeptly encapsulate the advancement and intensification of auditory attributes. The MFCCs, along with their first-order derivatives (deltas), and their second-order derivatives (double deltas) have a structural resemblance to the Red-Green-Blue (RGB) channels found in images. After they are resized, this structural similarity enables the representation of audio signals in the format of $3 \times 224 \times 224$, which is specifically designed to suit the DeiT model. This format optimization enhances the DeiT model's capability to detect VHD signals from audio data.

### B. CONVOLUTION MODULE

Relying solely on existing image-based models may not fully encompass the diverse range of audio intricacies. In order

to address this gap, we suggest the implementation of a hybrid convolution-transfer learning model that integrates a convolutional block with the SE attention mechanism.

The convolutional block, responsible for processing the MFCCs, is formally described as:

$$C_{\text{output}} = \text{Conv}(MFCC) \tag{4}$$

The module consists of three separate convolutional branches, each utilizing a kernel size of $3 \times 3$. The primary goal of the design is to extract channel features that are specific, while also preserving the original dimensions of the spectrogram. Additionally, the design attempts to improve the representation of features, even when the available data is low, without compromising computational efficiency.

The subsequent stage of the neural network architecture, known as the SE block, further enhances the results obtained by the convolution module:

$$SE_{\text{output}} = SE(C_{\text{output}}) \tag{5}$$

As indicated by Equation 5, this block adaptively tailors channel dependencies and accentuates key features, in accordance with the findings of [60].

### C. DeiT-BASED CLASSIFIER

The Vision Transformer (ViT) adapts the transformer architecture, initially developed for natural language processing (NLP) tasks [61], to effectively process image data. The power of DeiT resides in its ability to effectively merge data efficiency with the transformer model's proficient pattern recognition skills. Precision in distinguishing tiny heart sounds is of utmost importance in jobs such as VHD identification. In contrast to traditional transfer learning methods that largely depend on extensive labeled datasets, DeiT employs a distinctive distillation token to facilitate efficient learning even in scenarios with low data availability. This methodology offers the potential for expedited convergence and improved performance, hence conferring a competitive edge over conventional models such as CNN [57]. The main operation of the ViT is described as post-processing by the self-attention mechanism module:

$$ViT_{\text{output}} = ViT(SE_{\text{output}}) \tag{6}$$

The inclusion of a ''classification token'' is a crucial component in conventional ViT architectures. Nevertheless, the absence of spatially-focused layers in Vision Transformers necessitates a substantial amount of pre-training in order to achieve comparable performance to CNN. In response to this particular difficulty, DeiT proposes the implementation of a distillation token. This token serves the purpose of simulating label predictions, under the guidance of a mentor model.

### D. LOSS FUNCTION

The culmination of the process is the evaluation of loss function. Using the output from the Vision Transformer and the actual labels, the loss is computed as:

$$L = \mathcal{L}(ViT_{\text{output}}, y) \tag{7}$$

where $y$ denotes the true label.

In essence, the end-to-end relationship, commencing from the raw audio signal and culminating in the computed loss, is encapsulated by:

$$L = \mathcal{L}(ViT(SE(Conv(F(S)))), y) \tag{8}$$

This relationship offers a comprehensive perspective on the process of data travel inside the system, outlining the progression from raw audio to the ultimate assessment of loss.

**TABLE 2. Detail of the HSM database.**

| Cardiac Disorder | No. of Sample |
|---|---|
| AS | 200 |
| MR | 200 |
| MS | 200 |
| MVP | 200 |
| N | 200 |
| Total | 1,000 |

## III. EXPERIMENT SETUP

### A. DATABASE

The Heart Sound Murmur (HSM) database [62] was utilized in this investigation. The database consists of 1000 samples of phonocardiogram (PCG), which have been formatted as .wav audio files. The recordings are characterized by a single-channel configuration, featuring a bit-depth of 16 bits per sample and a sampling rate of 8000 Hz. The collected samples encompass five distinct categories, namely Aortic Stenosis (AS), Mitral Regurgitation (MR), Mitral Stenosis (MS), Mitral Valve Prolapse (MVP), and Normal (N). Further details on the datasets can be found in Table 2.

### B. FEATURE EXTRACTION

In this paper, the MFCCs were utilized as the principal magnitude characteristic for the input of our transformer model. The computation of these coefficients was performed using a frame length of 20 ms, with a 50% overlap, and the application of a Hamming window to each frame. This process yielded a 38-dimensional MFCC. In order to capture the temporal dynamics between audio frames, both delta and double-delta coefficients of MFCC were computed. According to the study conducted by [47], a predetermined segment duration of 2 seconds was employed. This duration was determined to capture a greater quantity of information compared to segments of 1 second or 1.5 seconds, based on empirical experimentation. The MFCC data, together with its delta and double delta, was reshaped into dimensions of $3 \times 224\text{X}224$ (representing channels, height, and width). This reshaping was done to ensure compliance with the pre-trained DeiT model, which is specifically designed for image datasets.

**TABLE 3.** Model parameters of Conv-DeiT.

| Parameters | Conv-DeiT |
|---|---|
| Batch Size | 64 |
| Learning Rate | 0.001 |
| Epochs | 100 |
| Optimizer | ADAM |
| Loss | Cross-entropy |
| Output Activation | SoftMax |
| Hidden Layers | 128-5 |
| Hidden Layer Activation | ReLU |

### C. NETWORK TRAINING

The models for evaluation were constructed via the PyTorch v1.10.1 framework and conducted training on an NVIDIA RTX3090 GPU equipped with 24 GB of RAM. The training parameters are presented in Table 3.

A comprehensive set of experiments was conducted to assess the effects of various factors, including different learning rates and batch sizes. Nevertheless, these modifications did not enhance the detection performance. This indicates that adjusting parameters did not yield positive results for our Conv-DeiT approach or other similar transfer learning classifiers. The selected values showed a desirable balance between computing requirements and the effectiveness of the model.

### D. PERFORMANCE CRITERIA

For classifier performance evaluation, prevalent matrices were adopted as suggested in [47]: precision (Pr), recall (Re), F1-score (F1), and accuracy (Acc). These matrices are determined by:

$$Pr = \frac{TP}{TP + FP} \tag{9}$$

$$Re = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = 2 \times \frac{Pr \times Re}{Pr + Re} \tag{11}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

Here, $TP$, $TN$, $FP$, and $FN$ represent true positive, true negative, false positive, and false negative values respectively.

## IV. RESULTS AND DISCUSSIONS

### A. RESULT BASED ON THE PROPOSED METHODS

In this study, we analyze the performance of DeiT transfer learning by the utilization of various experimental approaches, which include:

- **DeiT-Small:** The present methodology employs the DeiT-Small model for the purpose of transfer learning, whereby features are extracted directly without the utilization of the Convolutional block. The DeiT-Small model is designed with 384 embedding dimensions, which determines the size of the hidden vector representations in the transformer model. Additionally, it employs 12 attention heads to effectively identify various patterns and correlations within the input data. The parameters for the DeiT-Small model is outlined in Table 3.
- **DeiT-Base:** The approach described in this study utilizes the DeiT-Base model for transfer learning by directly extracting features, similar to the DeiT-Small method, without making any modifications to the Convolutional block. The DeiT-Base model is notable for its utilization of 768 embedding dimensions and 6 attention heads. The parameters for the DeiT-Base model are outlined in Table 3.
- **Conv-DeiT w/o att:** The proposed approach incorporates the DeiT model-Base and the Conv block for transfer learning, while dropping the SE attention mechanism. The Conv-DeiT model without attention has the same configuration parameters as the DeiT-Base model.
- **Conv-DeiT:** This framework embraces the proposed structure depicted in Fig 1. The optimal parameters for the Conv-DeiT method are outlined in Table 3.

A comprehensive analysis of the performances of DeiT-small, DeiT-base, Conv-DeiT w/o att, and Conv-DeiT is presented in Table 4. The comparison is conducted throughout all five folds and encompasses many performance indicators. This complete perspective facilitates a thorough comprehension of the outcomes.

As depicted in Table 4, a comparative analysis was conducted on two variations of DeiT-based transfer learning techniques, distinguished by their respective parameter configurations. The results indicate that DeiT-base, which has 768 embedding dimensions and 6 heads, exhibited superior performance compared to DeiT-small, which has 384 embedding dimensions and 12 heads. The enhanced detection performance of DeiT-base can be attributed to its more intricate and resilient embedding representation. Therefore, the DeiT-base model is selected as the reference point for subsequent comparisons, either in conjunction with the Conv block or the Conv block with the SE-attention mechanism.

The Conv-DeiT approach, when implemented without attention, demonstrates greater performance compared to the DeiT method. The inclusion of the Conv block appears to play a significant role in enhancing the performance, as it offers more discernible information compared to the DeiT technique in the absence of the Conv block. As a result, the utilization of the Conv block in combination with the SE-attention mechanism leads to a more advanced embedding representation. Adding the attention mechanism to Conv-DeiT may allow it to outperform its version without attention.

After that, we proceeded to examine the statistical significance of the embedding features obtained from the flattened layers of DeiT-Small, DeiT-Base, Conv-DeiT w/o att, and Conv-DeiT. We utilised the Multivariate Analysis of Variance (MANOVA) technique [63] to assess the participant's capacity to differentiate among several categories of VHD.

**TABLE 4.** Classification results for different models.

| Classification scheme | Class | Pr (%) | Re (%) | F1(%) | Acc (%) |
|---|---|---|---|---|---|
| DeiT-Small | N | 100.00 | 97.93 | 98.95 | 99.60 |
| | MVP | 91.54 | 89.51 | 90.38 | 96.10 |
| | MR | 94.24 | 86.93 | 90.35 | 96.80 |
| | AS | 90.77 | 95.92 | 93.16 | 97.30 |
| | MS | 91.06 | 96.83 | 93.72 | 97.40 |
| | Avg | $93.52 \pm 3.87$ | $93.43 \pm 4.89$ | $93.31 \pm 3.51$ | $97.44 \pm 1.31$ |
| DeiT-Base | N | 97.47 | 99.01 | 98.22 | 99.30 |
| | MVP | 92.03 | 92.22 | 92.08 | 96.80 |
| | MR | 95.65 | 90.59 | 92.99 | 97.60 |
| | AS | 96.91 | 95.60 | 96.22 | 98.60 |
| | MS | 90.79 | 94.90 | 92.71 | 97.10 |
| | Avg | $94.57 \pm 2.99$ | $94.46 \pm 3.25$ | $94.44 \pm 2.65$ | $97.88 \pm 1.05$ |
| Conv-DeiT w/o att | N | 99.41 | 99.53 | 99.47 | 99.80 |
| | MVP | 91.42 | 93.46 | 92.24 | 96.80 |
| | MR | 95.32 | 92.46 | 93.79 | 98.00 |
| | AS | 97.19 | 97.48 | 97.30 | 98.90 |
| | MS | 94.02 | 94.74 | 94.21 | 97.70 |
| | Avg | $95.47 \pm 3.04$ | $95.53 \pm 2.92$ | $95.40 \pm 2.92$ | $98.24 \pm 1.15$ |
| Conv-DeiT | N | 100.00 | 100.00 | 100.00 | 100.00 |
| | MVP | 97.13 | 98.49 | 97.80 | 99.10 |
| | MR | 97.50 | 97.59 | 97.52 | 99.20 |
| | AS | 99.02 | 98.67 | 98.83 | 99.50 |
| | MS | 98.96 | 97.93 | 98.42 | 99.40 |
| | Avg | $98.52 \pm 1.18$ | $98.54 \pm 0.92$ | $98.51 \pm 0.98$ | $99.44 \pm 0.35$ |

**TABLE 5.** Statistical significance test of DeiT-Small, DeiT-Base, Conv-DeiT w/o att, and Conv-DeiT using MANOVA ($p$-value $< 0.05$).

| Models | Wilks' $\Lambda$ | Pillai's trace | Hotelling-Lawley trace | F-Value |
|---|---|---|---|---|
| DeTi-Small | 0.91 | 0.09 | 0.09 | 9.38 |
| DeTi-Base | 0.87 | 0.13 | 0.15 | 14.93 |
| Conv-DeiT w/o att | 0.82 | 0.18 | 0.22 | 21.55 |
| Conv-DeiT | 0.81 | 0.19 | 0.24 | 23.54 |

The Wilk's $\Lambda$ metric, Pillai's trace, Hotelling-Lawley trace, and the F-value were essential in this assessment. Lower values of Wilk's $\Lambda$ suggest greater statistical significance in distinguishing across categories of VHD. In contrast, larger values of the Pillai's trace, Hotelling-Lawley trace, and F-value indicate a higher level of statistical significance in distinguishing across VHD groups. The embedding features utilized in this investigation were derived from the initial experimental iteration. For this analysis, we used t-distributed Stochastic Neighbor Embedding (t-SNE) [64] to reduce the dimensionality of embedding features, crucial for visualizing and understanding the data. This step was vital to assess participants' ability to differentiate between VHD categories effectively. The statistical significance of the models DeiT-Small, DeiT-Base, Conv-DeiT w/o att, and Conv-DeiT is presented in Table 5. Our results underscore the Conv-DeiT model's distinct advantage in multiple statistical metrics, including Wilk's Lambda ($\Lambda$), Pillai's trace, Hotelling-Lawley trace, and the F-value. The model achieved the lowest Wilk's Lambda at 0.81 and the highest F-value at 23.54, indicating robust discriminative power in classifying various VHD categories, as detailed in Table 5. The Conv-DeiT model also recorded the highest Pillai's trace at 0.19 and Hotelling-Lawley trace at 0.24, further confirming its superior performance over the other models evaluated. These higher values reflect the model's enhanced sensitivity and accuracy in detecting differences among VHD groups. The integration of Squeeze-and-Excitation (SE) attention within the Conv-DeiT framework is instrumental to this performance, dynamically recalibrating channel-wise feature responses, which significantly improves the model's focus on relevant features for VHD detection. This strategic fusion of the DeiT architecture's global contextual awareness with SE attention's channel-specific refinement leads to precise and discerning feature representations. The Conv-DeiT model's superior statistical measures demonstrate the efficacy of this approach, marking a significant advancement in deep learning for medical imaging and specifically in the complex task of VHD classification. An in-depth technical exposition on the Conv-DeiT model's architecture and the functional integration of SE attention, which is foundational to the enhanced performance, is available in the supplementary materials.

In order to examine the distributions among VHDs and visualize the discriminatory information based on the Conv-Deit feature representation for VHD classification, we employed t-SNE, a well-known technique for dimensionality reduction. In this case, the outcome of the initial fold was selected. Fig. 2 presents the graphical representation of the flattened features obtained from the Conv-Deit model that underwent training.

As depicted in Fig 2(a), the data distributions of distinct groups utilizing unprocessed speech samples exhibited substantial overlap. This posed a difficulty in distinguishing

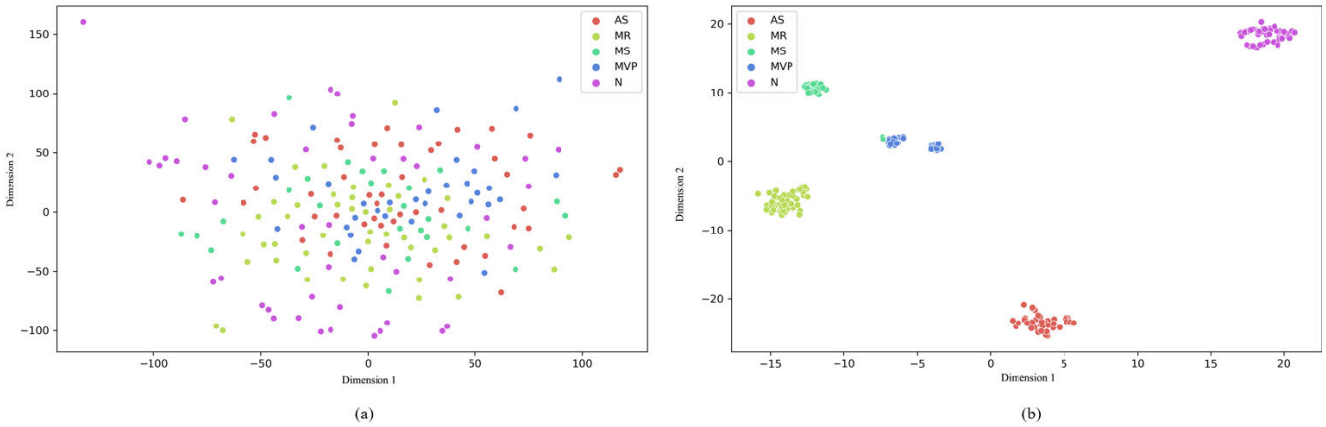



**FIGURE 2.** Visual distributions of different features based on t-SNE. (a) raw speech signals, (b) flatten features derived from Conv-DeiT model.

**TABLE 6.** Comparison with some known systems.

| Reference | Feature extraction | Classifier | Accuracy (%) |
|---|---|---|---|
| [5] | MFCC | KNN | 84.30 |
| [5] | MFCC | MLP | 80.90 |
| [5] | MFCC | SVM | 86.10 |
| [5] | MFCC | DNN | 89.10 |
| [5] | MFCCs based on EMD | KNN | 88.40 |
| [5] | MFCCs based on EMD | MLP | 88.80 |
| [5] | MFCCs based on EMD | SVM | 96.20 |
| [5] | MFCCs based on EMD | DNN | 98.90 |
| [55] | FDPCT | VGG16-based transfer learning | 94.00 |
| [55] | FDPCT | ResNet-50-based transfer learning | 98.00 |
| [55] | FDPCT | Deep CNN | 99.48 |
| [30] | Chirplet transform | MCC | 98.33 |
| [47] | log-mel spectrogram | LSTM | 95.00 |
| [47] | log-mel spectrogram | CNN | 99.67 |
| [48] | LPCC, MFCC | SFF-HLSTM | 99.10 |
| [48] | LPCC, MFCC | PFF-HLSTM | 98.71 |
| [66] | Tunable Q-factor wavelet transform | Deep Wavelet | 98.48 |
| **Ours** | **MFCC** | **DeiT-based transfer learning** | **97.88** |
| **Ours** | **MFCC** | **Conv-DeiT-based transfer learning** | **99.44** |

among various forms of VHD. However, it is evident from Fig 2(b) that the feature derived from the Conv-Deit model, after training, exhibited enhanced performance in comparison to utilizing unprocessed voice samples. The object had distinct outlines and shorter distances among different classes, suggesting its efficacy in discerning among various types of VHD. The findings of this study indicate that the spatial-temporal characteristic derived from the Conv-Deit model may hold significant value in the detection of VHD.

### B. COMPARISON WITH SOME KNOWN SYSTEMS

In this subsection, we evaluate the effectiveness of our suggested approaches by comparing them to established systems. As emphasized in the opening, conversations may circumvent specific frameworks when their empirical foundations deviate from our prescribed repository. The primary emphasis of our study is on the data obtained from the HSM database, which is consistent with the experimental framework in which we have established. Table 1 presents a comparison of the results obtained from different well-known systems in relation to our proposed technique.

As shown in Table 6, it is apparent that Conv-DeiT demonstrates superior performance compared to other established systems when evaluated on the HSM database. Nevertheless, the performance of the aforementioned model did not exceed that of the classifier based on Convolutional Neural Networks (CNN) utilizing the log-mel spectrogram feature, as reported by [47]. One plausible explanation may lie in the process of self-selection of training and testing datasets. However, the outcomes of our study are based on a five-fold cross-validation approach, which enhances the reliability of the classification performance of the offered approaches. Furthermore, it is worth noting that Conv-DeiT did not achieve superior performance compared to the deep CNN classifier while utilizing the FDPCT feature. The aforementioned statement highlights the importance of FDPCT's capacity to effectively analyze non-stationary signals, specifically those found in PCG signals. The combination of FDPCT with deep learning has been found to improve the accuracy of

VHD identification. While our Conv-DeiT model, designed to enhance DeiT, showed a marginal performance drop of 0.04%, it still presents a viable approach for VHD detection.

## V. CONCLUSION

In this paper, we propose identifying VHD by employing transfer learning methods that take advantage of pre-trained transformer models based on image data. The DeiT model, initially pre-trained on image datasets, was harnessed for its inherent capabilities. This strategy achieved a notable overall accuracy of 97.44% in the classification of heart sounds. Subsequent enhancements led to the proposal of the Conv-DeiT approach, a hybrid architecture that integrates a convolutional block, an SE attention mechanism, and the DeiT process. This method exhibited superior performance compared to the standalone DeiT-based transfer learning, reaching an exceptional overall performance of 99.44%. The study has indicated the potential of DeiT-based transfer learning and the efficacy of using models pre-trained on a distinct modality, such as images, for classifying heart sounds. Moreover, our hybrid Conv-DeiT method, which combines the convolutional block and the SE-attention mechanism has demonstrated significant advantages in this context.

In future work, we aim to investigate other attention mechanisms to further refine our proposed methods. We also plan to incorporate multi-scale convolutional neural networks [35] and phase information [49], [65] as supplementary data to enhance our methodologies.

## REFERENCES


[1] J. S. Aluru, A. Barsouk, K. Saginala, P. Rawla, and A. Barsouk, "Valvular heart disease epidemiology," *Med. Sci.*, vol. 10, no. 2, p. 32, Jun. 2022.

[2] C. W. Tsao et al., "Heart disease and stroke statistics—2022 update: A report from the American Heart Association," *Circulation*, vol. 145, no. 8, pp. e153–e639, Jan. 2022.

[3] J. Chen, S. Sun, L.-B. Zhang, B. Yang, and W. Wang, "Compressed sensing framework for heart sound acquisition in Internet of Medical Things," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 2000–2009, Mar. 2022.

[4] A. Quinn, J. Kaminsky, A. Adler, S. Eisner, and R. Ovitsh, "Cardiac auscultation lab using a heart sounds auscultation simulation manikin," *MedEdPORTAL*, vol. 15, p. 10839, Oct. 2019.

[5] Ö. Arslan and M. Karhan, "Effect of Hilbert–Huang transform on classification of PCG signals using machine learning," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 10, pp. 9915–9925, Nov. 2022.

[6] V. Arora, R. Leekha, R. Singh, and I. Chana, "Heart sound classification using machine learning and phonocardiogram," *Modern Phys. Lett. B*, vol. 33, no. 26, Sep. 2019, Art. no. 1950321.

[7] M. G. M. Milani, P. E. Abas, L. C. De Silva, and N. D. Nanayakkara, "Abnormal heart sound classification using phonocardiography signals," *Smart Health*, vol. 21, Jul. 2021, Art. no. 100194.

[8] J. S. Chorba et al., "Deep learning algorithm for automated cardiac murmur detection via a digital stethoscope platform," *J. Amer. Heart Assoc.*, vol. 10, no. 9, May 2021, Art. no. e019905.

[9] Y. Al-Issa and A. M. Alqudah, "A lightweight hybrid deep learning system for cardiac valvular disease classification," *Sci. Rep.*, vol. 12, no. 1, p. 14297, Aug. 2022.

[10] P. Jyothi and G. Pradeepini, "Review on cardiac arrhythmia through segmentation approaches in deep learning," in *Proc. Int. Conf. Intell. Smart Comput. Data Anal.*, Singapore, 2021, pp. 139–147.

[11] A. F. Gündüz and F. Talu, "PCG frame classification by classical machine learning methods using spectral features and MFCC based features," *Eur. J. Sci. Technol.*, vol. 42, pp. 77–82, Oct. 2022.

[12] M. Nassralla, Z. E. Zein, and H. Hajj, "Classification of normal and abnormal heart sounds," in *Proc. 4th Int. Conf. Adv. Biomed. Eng. (ICABME)*, Oct. 2017, pp. 1–4.

[13] M. Zabihi, A. B. Rad, S. Kiranyaz, M. Gabbouj, and A. K. Katsaggelos, "Heart sound anomaly and quality detection using ensemble of neural networks without segmentation," in *Proc. Comput. Cardiol. Conf. (CinC)*, vol. 43, Sep. 2016, pp. 613–616.

[14] T.-E. Chen, S.-I. Yang, L.-T. Ho, K.-H. Tsai, Y.-H. Chen, Y.-F. Chang, Y.-H. Lai, S.-S. Wang, Y. Tsao, and C.-C. Wu, "S1 and S2 heart sound recognition using deep neural networks," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 2, pp. 372–380, Feb. 2017.

[15] X. Bao, Y. Xu, and E. N. Kamavuako, "The effect of signal duration on the classification of heart sounds: A deep learning approach," *Sensors*, vol. 22, no. 6, p. 2261, Mar. 2022.

[16] W. Zeng, Z. Lin, C. Yuan, Q. Wang, F. Liu, and Y. Wang, "Detection of heart valve disorders from PCG signals using TQWT, FA-MVEMD, Shannon energy envelope and deterministic learning," *Artif. Intell. Rev.*, vol. 54, no. 8, pp. 6063–6100, Dec. 2021.

[17] Z. Abduh, E. A. Nehary, M. A. Wahed, and Y. M. Kadah, "Classification of heart sounds using fractional Fourier transform based mel-frequency spectral coefficients and traditional classifiers," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101788.

[18] F. Beritelli, G. Capizzi, G. Lo Sciuto, C. Napoli, and F. Scaglione, "Automatic heart activity diagnosis based on Gram polynomials and probabilistic neural networks," *Biomed. Eng. Lett.*, vol. 8, no. 1, pp. 77–85, Feb. 2018.

[19] R. Touahria, A. Hacine-Gharbi, and P. Ravier, "Discrete wavelet based features for PCG signal classification using hidden Markov models," in *Proc. 10th Int. Conf. Pattern Recognit. Appl. Methods*, 2021, pp. 334–340.

[20] M. Tschannen, T. Kramer, G. Marti, M. Heinzmann, and T. Wiatowski, "Heart sound classification using deep structured features," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2016, pp. 565–568.

[21] Y. Chen, S. Wei, and Y. Zhang, "Classification of heart sounds based on the combination of the modified frequency wavelet transform and convolutional neural network," *Med. Biol. Eng. Comput.*, vol. 58, no. 9, pp. 2039–2047, Sep. 2020.

[22] H. Tang, Z. Dai, Y. Jiang, T. Li, and C. Liu, "PCG classification using multidomain features and SVM classifier," *BioMed Res. Int.*, vol. 2018, pp. 1–14, Jul. 2018.

[23] M. U. Khan, S. Aziz, K. Iqtidar, G. F. Zaher, S. Alghamdi, and M. Gull, "A two-stage classification model integrating feature fusion for coronary artery disease detection and classification," *Multimedia Tools Appl.*, vol. 81, no. 10, pp. 13661–13690, Apr. 2022.

[24] T. Indu, A. Prakash, S. R. Chandran, N. Babu, and R. Soorya, "Comparison of different machine learning algorithms for cardiac auscultation," in *Proc. IEEE Int. Conf. Signal Process., Informat., Commun. Energy Syst. (SPICES)*, vol. 1, Mar. 2022, pp. 113–117.

[25] J. J. G. Ortiz, C. P. Phoo, and J. Wiens, "Heart sound classification based on temporal alignment techniques," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2016, pp. 589–592.

[26] W. Yang, J. Xu, J. Xiang, Z. Yan, H. Zhou, B. Wen, H. Kong, R. Zhu, and W. Li, "Diagnosis of cardiac abnormalities based on phonocardiogram using a novel fuzzy matching feature extraction method," *BMC Med. Informat. Decis. Making*, vol. 22, no. 1, pp. 1–13, Sep. 2022.

[27] D. S. Panah, A. Hines, and S. Mckeever, "Exploring composite dataset biases for heart sound classification," in *Proc. 28th Irish Conf. Artif. Intell. Cogn. Sci.*, Dec. 2020, pp. 145–156.

[28] S. K. Ghosh, R. N. Ponnalagu, R. K. Tripathy, and U. R. Acharya, "Automated detection of heart valve diseases using chirplet transform and multiclass composite classifier with PCG signals," *Comput. Biol. Med.*, vol. 118, Mar. 2020, Art. no. 103632.

[29] S. Das, S. Pal, and M. Mitra, "Supervised model for cochleagram feature based fundamental heart sound identification," *Biomed. Signal Process. Control*, vol. 52, pp. 32–40, Jul. 2019.

[30] S. K. Ghosh, R. K. Tripathy, R. N. Ponnalagu, and R. B. Pachori, "Automated detection of heart valve disorders from the PCG signal using time-frequency magnitude and phase features," *IEEE Sensors Lett.*, vol. 3, no. 12, pp. 1–4, Dec. 2019.

[31] A. F. Quiceno-Manrique, J. I. Godino-Llorente, M. Blanco-Velasco, and G. Castellanos-Dominguez, "Selection of dynamic features based on time-frequency representations for heart murmur detection from phonocardiographic signals," *Ann. Biomed. Eng.*, vol. 38, no. 1, pp. 118–137, Jan. 2010.

[32] U. Riaz, S. Aziz, M. U. Khan, S. A. A. Zaidi, M. Ukasha, and A. Rashid, "A novel embedded system design for the detection and classification of cardiac disorders," *Comput. Intell.*, vol. 37, no. 4, pp. 1844–1864, Nov. 2021.

[33] O. El Badlaoui, A. Benba, and A. Hammouch, "Novel PCG analysis method for discriminating between abnormal and normal heart sounds," *IRBM*, vol. 41, no. 4, pp. 223–228, Aug. 2020.

[34] M. S. Ahmad, J. Mir, M. O. Ullah, M. L. U. R. Shahid, and M. A. Syed, "An efficient heart murmur recognition and cardiovascular disorders classification system," *Australas. Phys. Eng. Sci. Med.*, vol. 42, no. 3, pp. 733–743, Sep. 2019.

[35] W. Pathonsuwan, K. Phapatanaburi, P. Buayai, T. Jumphoo, P. Anchuen, M. Uthansakul, and P. Uthansakul, "RS-MSConvNet: A novel end-to-end pathological voice detection model," *IEEE Access*, vol. 10, pp. 120450–120461, 2022.

[36] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[37] S. Latif, M. Usman, R. Rana, and J. Qadir, "Phonocardiographic sensing using deep learning for abnormal heartbeat detection," *IEEE Sensors J.*, vol. 18, no. 22, pp. 9393–9400, Nov. 2018.

[38] M. Sotaquirá, D. Alvear, and M. Mondragón, "Phonocardiogram classification using deep neural networks and weighted probability comparisons," *J. Med. Eng. Technol.*, vol. 42, no. 7, pp. 510–517, Oct. 2018.

[39] B. Bozkurt, I. Germanakis, and Y. Stylianou, "A study of time-frequency features for CNN-based automatic heart sound classification for pathology detection," *Comput. Biol. Med.*, vol. 100, pp. 132–143, Sep. 2018.

[40] C. Potes, S. Parvaneh, A. Rahman, and B. Conroy, "Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds," in *Proc. Comput. Cardiol. Conf. (CinC)*, Sep. 2016, pp. 621–624.

[41] J. M.-T. Wu, M.-H. Tsai, Y. Z. Huang, S. H. Islam, M. M. Hassan, A. Alelaiwi, and G. Fortino, "Applying an ensemble convolutional neural network with Savitzky–Golay filter to construct a phonocardiogram prediction model," *Appl. Soft Comput.*, vol. 78, pp. 29–40, May 2019.

[42] S. L. Oh, V. Jahmunah, C. P. Ooi, R.-S. Tan, E. J. Ciaccio, T. Yamakawa, M. Tanabe, M. Kobayashi, and U. R. Acharya, "Classification of heart sound signals using a novel deep WaveNet model," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105604.

[43] J. P. Dominguez-Morales, A. F. Jimenez-Fernandez, M. J. Dominguez-Morales, and G. Jimenez-Moreno, "Deep neural networks for the recognition and classification of heart murmurs using neuromorphic auditory sensors," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 1, pp. 24–34, Feb. 2018.

[44] Y. Xu, B. Xiao, X. Bi, W. Li, J. Zhang, and X. Ma, "Pay more attention with fewer parameters: A novel 1-D convolutional neural network for heart sounds classification," in *Proc. Comput. Cardiol. Conf. (CinC)*, vol. 45, Sep. 2018, pp. 1–4.

[45] B. Xiao, Y. Xu, X. Bi, W. Li, Z. Ma, J. Zhang, and X. Ma, "Follow the sound of children's heart: A deep-learning-based computer-aided pediatric CHDs diagnosis system," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1994–2004, Mar. 2020.

[46] F. Li, M. Liu, Y. Zhao, L. Kong, L. Dong, X. Liu, and M. Hui, "Feature extraction and classification of heart sound using 1D convolutional neural networks," *EURASIP J. Adv. Signal Process.*, vol. 2019, no. 1, p. 111, Dec. 2019.

[47] M. T. Nguyen, W. W. Lin, and J. H. Huang, "Heart sound classification using deep learning techniques based on log-mel spectrogram," *Circuits, Syst., Signal Process.*, vol. 42, no. 1, pp. 344–360, Jan. 2023.

[48] S. Das, D. Jyotishi, and S. Dandapat, "Heart valve diseases detection based on feature-fusion and hierarchical LSTM network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[49] K. Phapatanaburi, W. Pathonsuwan, L. Wang, P. Anchuen, T. Jumphoo, P. Buayai, M. Uthansakul, and P. Uthansakul, "Whispered speech detection using glottal flow-based features," *Symmetry*, vol. 14, no. 4, p. 777, Apr. 2022.

[50] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: A literature review," *BMC Med. Imag.*, vol. 22, no. 1, p. 69, Dec. 2022.

[51] Z. Wan, R. Yang, M. Huang, N. Zeng, and X. Liu, "A review on transfer learning in EEG signal analysis," *Neurocomputing*, vol. 421, pp. 1–14, Jan. 2021.

[52] T. Alafif, M. Boulares, A. Barnawi, T. Alafif, H. Althobaiti, and A. Alferaidi, "Normal and abnormal heart rates recognition using transfer learning," in *Proc. 12th Int. Conf. Knowl. Syst. Eng. (KSE)*, Nov. 2020, pp. 275–280.

[53] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1251–1258.

[54] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014.

[55] J. Karhade, S. Dash, S. K. Ghosh, D. K. Dash, and R. K. Tripathy, "Time-frequency-domain deep learning framework for the automated detection of heart valve disorders using PCG signals," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–11, 2022.

[56] C. Liu et al., "An open access database for the evaluation of heart sound algorithms," *Physiolog. Meas.*, vol. 37, no. 12, pp. 2181–2213, Dec. 2016.

[57] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jgou, "Training data-efficient image transformers & distillation through attention," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[58] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Salt Lake City, UT, USA, 2001, pp. 73–76.

[59] L. Rabiner and R. Schafer, *Theory and Applications of Digital Speech Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 2010.

[60] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.

[61] 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 × 16 words: Transformers for image recognition at scale," in *Proc. 9th Int. Conf. Learn. Represent.*, 2021.

[62] Yaseen, G.-Y. Son, and S. Kwon, "Classification of heart sound signal using multiple features," *Appl. Sci.*, vol. 8, no. 12, p. 2344, Nov. 2018.

[63] W. J. Krzanowski, *Principles of Multivariate Analysis: A Users Perspective*. Oxford, U.K.: Clarendon, 1988.

[64] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007.

[65] K. Phapatanaburi, P. Buayai, M. Kupimai, and T. Yodrot, "Linear prediction residual-based constant-Q cepstral coefficients for replay attack detection," in *Proc. 8th Int. Electr. Eng. Congr. (iEECON)*, Chiang Mai, Thailand, Mar. 2020, pp. 1–4.

[66] W. Zeng, J. Yuan, C. Yuan, Q. Wang, F. Liu, and Y. Wang, "A new approach for the detection of abnormal heart sound signals using TQWT, VMD and neural networks," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1613–1647, Mar. 2021.

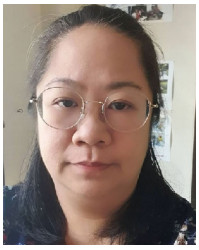
[67] P. Langley and A. Murray, "Abnormal heart sounds detected from short duration unsegmented phonocardiograms by wavelet entropy," in *Proc. Comput. Cardiol. Conf. (CinC)*, Vancouver, BC, Canada, Sep. 2016, pp. 545–548.

**TALIT JUMPHOO** received the B.E. degree in telecommunication and electronic engineering and the Ph.D. degree in telecommunication engineering from the Suranaree University of Technology, Thailand, in 2014 and 2022, respectively. He is currently a Postdoctoral Researcher with the School of Telecommunication Engineering, Institute of Engineering, Suranaree University of Technology. His research interests included biosignal processing, biomedical devices, brain–computer interface, and applied machine learning.

**KHOMDET PHAPATANABURI** received the B.E. degree in electronic and telecommunication engineering and the M.E. degree in electrical engineering from the Rajamangala University of Technology Thanyaburi (RMUTT), Thailand, in 2010 and 2012, respectively, and the Dr.Eng. degree in information science and control engineering from the Nagaoka University of Technology (NUT), Japan, in 2017. In 2018, he joined the Department of Telecommunication Engineering, Rajamangala University of Technology Isan, as a Lecturer, and became an Assistant Professor, in 2020. During his study in Japan, he received the Monbukagakusho (MEXT) Scholarship, from 2014 to 2017. His main research interest includes audio and brainwave classification. He is a Reviewer of several international journals, including IEEE SIGNAL PROCESSING LETTERS, *Computer Speech and Language*, *APSIPA Transactions on Signal and Information Processing*, and IEEE ACCESS.

**WONGSATHON PATHONSUWAN** received the B.E. degree in telecommunication engineering and the M.E. degree in telecommunication and computer engineering from the Suranaree University of Technology, Thailand, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in telecommunication and computer. His research interests include wireless communication, artificial intelligent (AI), machine learning (ML), and artificial neural networks (ANN).

**PATIKORN ANCHUEN** received the B.E. degree in telecommunication engineering from the King Mongkut's Institute of Technology Ladkrabang, Thailand, in 2014, and the M.E. degree in telecommunication engineering and the Ph.D. degree in telecommunication and computer engineering from the Suranaree University of Technology, Thailand, in 2017 and 2020, respectively. He is currently a Lecturer with the Office of Graduate Studies, Navaminda Kasatriyadhiraj Royal Air Force Academy, Thailand. His research interests include wireless communication, artificial intelligent (AI), machine learning (ML), artificial neural networks (ANN), deep reinforcement learning (DRL), genetic algorithm (GA), particle swarm optimization (PSO), mobile networks, quality of experience (QoE), and 5G communications.



**MONTHIPPA UTHANSAKUL** (Member, IEEE) received the B.E. degree in telecommunication engineering from the Suranaree University of Technology, Thailand, in 1997, the M.E. degree in electrical engineering from Chulalongkorn University, Thailand, in 1999, and the Ph.D. degree in information technology and electrical engineering from The University of Queensland, Australia, in 2007. She is currently an Associate Professor with the Telecommunication School, Suranaree University of Technology. Her research interests include wideband/narrowband smart antennas, automatic switch beam antenna, DOA finder, microwave components, application of smart antenna, and advance wireless communications. She received the Second Prize of Young Scientist Award from 16th International Conference on Microwaves, Radar, and Wireless Communications, Poland, in 2006.

**PEERAPONG UTHANSAKUL** (Member, IEEE) received the B.E. and M.E. degrees in electrical engineering from Chulalongkorn University, Bangkok, Thailand, in 1996 and 1998, respectively, and the Ph.D. degree in information technology and electrical engineering from The University of Queensland, Brisbane, QLD, Australia, in 2007. From 1998 to 2001, he was a telecommunication engineer with one of the leading telecommunication companies in Thailand. He is currently an Associate Professor and the Dean of the Research Department, Suranaree University of Technology, Nakhon Ratchasima, Thailand. He has more than 100 research publications and the author/coauthor of various books related to MIMO technologies. His research interests include green communications, wave propagation modeling, MIMO, massive MIMO, brain wave engineering, OFDM and advanced wireless communications, wireless sensor networks, embedded systems, the Internet of Things, and network security. He has won various national awards from the Government of Thailand for his contributions and motivation in the field of science and technology. Furthermore, he is an Editor of *Suranaree Journal of Science and Technology*; and other leading Thailand journals related to science and technology and wireless communications, Poland, in 2006.