Diagnosis of Peripheral Artery Disease Using Backflow Abnormalities in Proximal Recordings of Accelerometer Contact Microphone (ACM)

Arash Shokouhmand, *Student Member, IEEE*, Haoran Wen, *Member, IEEE*, Samiha Khan, Joseph A. Puma, Amisha Patel, Philip Green, Farrokh Ayazi, *Fellow, IEEE*, and Negar Tavassolian, *Senior Member, IEEE*

Abstract*—*Objective: **The development of an accurate, non-invasive method for the diagnosis of peripheral artery disease (PAD) from accelerometer contact microphone (ACM) recordings of the cardiac system.** ***Methods:*** **Mel frequency cepstral coefficients (MFCCs) are initially extracted from ACM recordings. The extracted MFCCs are then used to fine-tune a pre-trained ResNet50 network whose middle layers provide streams of high-level-of-abstraction coefficients (HLACs) which could provide information on blood pressure backflow caused by arterial obstructions in PAD patients. A vision transformer is finally integrated with the feature extraction layer to detect PAD, and stratify the severity level. This architecture is coined multi-stream-powered vision transformer (MSPViT). The performance of MSPViT is evaluated on 74 PAD and 21 healthy subjects.** ***Results:*** **Sensitivity, specificity, F1 score, and area under the curve (AUC) of 99.45%, 98.21%, 99.37%, and 0.99, respectively, are reported for the binary classification which ensures accurate detection of PAD. Furthermore, MSPViT suggests average sensitivity, specificity, F1 score, and AUC of 96.66%, 97.34%, 96.29%, and 0.96, respectively, for the classification of subjects into healthy, mild-PAD, and severe-PAD classes. The silhouette score is calculated to assess the separability of clusters formed for classes in the penultimate layer of MSPViT. An average silhouette score of 0.66 and 0.81 demonstrate excellent cluster separability in PAD detection and severity classification, respectively.** ***Conclusion:*** **The achieved performance suggests that the proximal ACM-driven framework can replace state-of-the-art techniques for PAD detection.** ***Significance:*** **This study presents a fundamental step towards prompt and accurate diagnosis of PAD and stratification of its severity level.**

***Index Terms*—Peripheral artery disease (PAD), accelerometer contact microphone (ACM), non-invasive recordings, multi-stream-powered, vision transformer.**

I. Introduction

PERIPHERAL artery disease (PAD), defined as the narrowing or obstruction of major systemic non-coronary arteries, currently affects 8.5 million people in the US [1]. The most common cause of PAD is the deposition of plaque in the inner layer of the arteries which is either asymptomatic, or accompanied by intermittent cramping pain in the leg known as claudication [2].

Asymptomatic PAD constitutes 20%–50% of the whole PAD population, inhibiting early-stage diagnosis and risk management [3]. Hence, the disease may progress into impaired blood flow to limbs, resulting in limb soreness, infection, and gangrene, and eventually ends up with limb amputation [4]. On the other hand, early detection of PAD allows for timely intervention such as supervised exercise, medications, and revascularization to prevent adverse outcomes [5].

The ankle/brachial index (ABI) is a commonly used non-invasive test for the detection of PAD. ABI is the ratio of the systolic blood pressure at the ankle to the systolic blood pressure at the arm, suggesting the severity level of PAD [6]. A low ABI (<0.90) is highly associated with the presence of occluded arteries and thus a powerful indicative of PAD [7]. Despite the ABI test being a convenient clinical practice, a reputable research conducted on 464 PAD patients in [8] reported a predictive sensitivity of 79% for the ABI test, which implies a low diagnostic performance. Angiography is the gold standard for PAD screening which offers direct visualization of the structure of an artery [9]. The procedure begins with injecting a radiographic contrast dye through a narrow, flexible catheter into the artery, which is followed by capturing X-ray images to view obstructed arterial regions with high sensitivity of 98% as reported by [10], [11]. The invasive nature of angiography however, causes local pain, puncture site, and discomfort for the patient [12]. The aforementioned shortcomings of the ABI test

Manuscript received 5 September 2022; revised 17 October 2022; accepted 29 October 2022. Date of publication 1 November 2022; date of current version 5 January 2023. This work was supported by the National Science Foundation (NSF) under Award 1855394. *(Corresponding author: Negar Tavassolian.)*

Arash Shokouhmand and Negar Tavassolian are with the Department of Electrical and Computer Engineering, Stevens Institute of Technology, Hoboken, NJ 07030 USA, and also with the Stanford School of Medicine at Stanford University, Stanford, CA 94305 USA (e-mail: ashokouh@stevens.edu; negar.tavassolian@stevens.edu).

Haoran Wen is with the StethX Microsystems Inc., Atlanta, GA 30308 USA (e-mail: haoran@stethx.com).

Samiha Khan is with the New York Institute of Technology College of Osteopathic Medicine, New York, NY 11545 USA (e-mail: skhan151@nyit.edu).

Joseph A. Puma, Amisha Patel, and Philip Green are with the Sorin Medical P.C., New York, NY 11207 USA (e-mail: jpuma@sorinmedicalny.com; ajp2001@caa.columbia.edu; pgreen@sorinmedicalny.com).

Farrokh Ayazi is with the StethX Microsystems Inc., Atlanta, GA 30308 USA, and also with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: farrokh.ayazi@ece.gatech.edu).

Digital Object Identifier 10.1109/JBHI.2022.3218595

and angiography provoke the need for a reliable, yet convenient screening method for PAD detection.

Recent advances in wearable sensors and machine learning techniques have established the basis for the non-invasive diagnosis of PAD. Two primary methods which are currently investigated at research level include oscillometry and photoplethysmography (PPG). In oscillometry, a blood pressure (BP) cuff is fully inflated around the subject's target limb [13], [14]. The superposition of the internal arterial pressure and the external pressure by the cuff is then measured in a gradual deflation using a pressure transducer placed in the cuff. The temporal evolution of the transducer recording is analyzed by a machine learning algorithm to identify the presence of PAD. The authors in [14] trained a multilayer perceptron model on the oscillometry data collected from 14 PAD patients and 19 healthy subjects. They reported an accuracy of 91.4%, sensitivity of 90.0%, and specificity of 92.1% for the diagnosis of PAD. These values imply a missed rate of 10% and a false detection rate of 7.9% for PAD detection, questioning the prediction stability of the oscillometry method. Furthermore, the cuff used in the oscillometry technique may cause inconvenience and pain for PAD patients with claudication in the calf region.

PPG is a non-invasive method of estimating blood flow variations in arteries, where an optical sensor continuously emits light with an excitation wavelength of 680–950 nm to estimate the instantaneous blood volume based on the reflected energy from the artery [15]. As proven by several studies, temporal and spectral characteristics of PPG signals recorded at toes [16], [17], [18], [19] and fingers [17], [20] can contribute to the diagnosis of PAD. According to the study conducted on 30 PAD patients in [17], it was demonstrated that the toe PPG could provide sensitivity and specificity of 0.95 and 0.65 for the diagnosis of PAD, respectively. As reported by [19], toe PPG recordings can be used with deep neural networks to achieve a sensitivity of 86.6% and specificity of 90.2% for PAD detection. However, PPG morphology can be affected by environmental lighting [16], as well as physiological characteristics such as age and health status [21], [22]. A common characteristic among the aforementioned methods is the use of either a BP cuff or a PPG sensor placed on the arterial site. Both methods fail to provide acceptable prediction power as reported by [13], [14], [16], [17], [18], [19], [20].

In this paper, we take advantage of the dynamics of blood pressure backflow in obstructed arteries, and establish a novel framework for the detection of PAD. The major contributions of this paper are as follows:

- This paper introduces a novel method for the diagnosis of PAD using cardiac cycle information collected by a high-precision accelerometer contact microphone. To this end, 95 subjects, including 74 PAD and 21 healthy subjects, were monitored. To the best of our knowledge, this is the first study that addresses the detection of PAD from blood pressure backflow patterns in cardiac cycles.
- A novel feature extraction technique based on transfer learning is introduced that extracts blood pressure backflow patterns and relates them to the presence of PAD and its severity level. The feature extraction approach is based on the use of high-level-of-abstraction feature maps derived from the middle layers of a pre-trained deep neural network.
- A modified vision transformer, coined multi-stream-powered vision transformer, is developed to detect PAD and classify its severity level. This neural network leverages the dynamics of ACM recordings in different frequency bands to recognize patterns associated with peripheral artery disease.

This paper is organized as follows: Section II provides background on the dynamics of the arterial tree, and elaborates their relation with PAD. In Section III, the experimental setup and methodology are described. Experimental results are presented in Section IV and comprehensively discussed in Section VI. The paper is concluded in Section V.

II. Background on Arterial Tree Dynamics

The arterial tree represents the branching system of arteries which are responsible for carrying oxygen-rich blood from the heart to other organs. According to the multi-branch transmission lines (TLs) model introduced by [23], the propagation of blood flow in an artery can be formulated by:

$$q_{\text{out}} = \frac{q_{\text{in}}^{(1-\Gamma)}}{e^{\gamma l} - \Gamma e^{-\gamma l}}, \tag{1}$$

where q_{in}, q_{out}, Γ, γ, and l denote the inlet blood flow (mL/s), outlet blood flow (mL/s), reflection coefficient, propagation constant (rad/cm), and arterial length (cm), respectively. For a fully obstructed artery, the left-hand side of (1) holds a zero value since there is no outlet blood flow. Hence, assuming $|q_{\text{in}}| > 1$,

$$q_{\text{in}}^{(1-\Gamma)} = 0, \tag{2}$$

which implies either $\Gamma \to +\infty$, or

$$e^{\gamma l} - \Gamma e^{-\gamma l} \to \pm\infty. \tag{3}$$

For a certain artery length of l, we can consider $e^{\gamma l}$ and $e^{-\gamma l}$ to be constant positive values. Hence, (3) results in $\Gamma \to \pm\infty$. As the reflection coefficient (Γ) holds positive values only, the term $\Gamma \to +\infty$ is acceptable. Terms (2) and (3) suggest that the more severe the obstruction, the larger the reflection of the blood flow. The reflection generates a resistive power towards the heart muscle pumping blood into the arterial tree. Consequently, the heart has to pump harder to force the blood through the obstructed arteries, causing the appearance of cardio-mechanical abnormalities in cardiac cycles [24]. In this study, we hypothesize that vibrational patterns associated with backflow in PAD patients emerge in the proximal recordings of cardiac cycles. To prove our hypothesis, we devise a pattern recognition framework to automatically differentiate healthy subjects from PAD patients based on the vibrational characteristics of their cardiac system.

III. Experimental Setup and Methodology

A. Data Collection and Study Protocol

In this study, seventy-four peripheral artery disease (PAD) patients including 35 males and 39 females were studied at the cardiac care unit of Sorin Medical P.C. Two interventional

TABLE I
DEMOGRAPHIC INFORMATION OF THE HEALTHY AND PAD PARTICIPANTS

Cohort	Number	Age	BMI (kg/m^2)
Healthy	21 (7 M, 14 F)	64.61 ± 16.43	29.15 ± 6.37
Mild PAD	29 (13 M, 16 F)	71.48 ± 12.05	29.20 ± 5.91
Severe PAD	45 (22 M, 23 F)	75.35 ± 9.08	29.24 ± 6.13

cardiologists and a general practitioner used the resting ABI as the screening and diagnostic test for PAD. ABI readings were categorized as abnormal (ABI < 0.90), borderline (0.91 < ABI < 0.99), normal (1.00 < ABI < 1.40), and non-compressible (ABI > 1.40) which is associated with arterial calcification in diabetes patients. The cardiologists categorized the pa-tients into 3 groups, 1) healthy (asymptomatic with normal ABI), 2) mild-PAD without intervention/revascularization (mild to moderate symptoms with either borderline, abnormal, or non-compressible ABI), and 3) severe-PAD with interven-tion/revascularization (severe symptoms with abnormal ABI when initial conservative treatments have not effectively reduced the symptoms). The mild-PAD and severe-PAD groups have average (±standard deviation) ages of 71.48 (±12.05) and 75.35 (±9.08) years, respectively. The control group consisting of 7 healthy males and 14 healthy females has average (±standard deviation) age of 64.61 (±16.43) years. Further demographic information on the subjects is summarized in Table I.

Fig. 1 illustrates the proposed framework for the detection of PAD. As shown in Fig. 1(a), the experimental setup consists of a ± 4 g sensitive accelerometer contact microphone (ACM) with micro-g resolution [25], [26] (obtained from StethX Mi-crosystems Inc., Atlanta, USA) attached to the chest wall along the third rib using medical-grade adhesive tape. Additionally, the epxerimental setup includes a three-lead ECG sensor node (ECG Development Kit; Shimmer Sensing, Dublin, Ireland) with the electrodes attached to the right arm (RA), left arm (LA), right leg (RL), and left leg (LL), as well as a photoplethysmogra-phy (PPG) sensor (Shimmer3 GSR; Shimmer Sensing, Dublin, Ireland) attached to the ear-lobe using an ear clip. The ACM sensor head has a small form factor of 27 mm × 15 mm × 2.5 mm which is comparable to the size of a dime as depicted in Fig. 1(b). This device is a low-noise accelerometer with a wide operational bandwidth of 0–10 kHz, allowing for recording heartbeat-induced sounds and vibrations on the chest wall. This device is not sensitive to airborne emission sounds, making it a robust phonocardiogram sensor against acoustic ambient noise. The data is digitized using a 24-bit analog-to-digital converter integrated in the sensor head, and transferred to a computer in real-time through a data transfer cable connected to the data acquisition (DAQ) module, as shown in Fig. 1(c). The subjects were seated at rest on a chair for a period of five minutes, followed by five minutes of ACM measurements at a sampling rate of 22.33 kHz, still at a seated position. The patient experimental protocol was approved by the Institutional Review Board of Stevens Institute of Technology under protocol number 2022-044 (N). It is to be noted that ECG and PPG data are not used in this specific study. In the proposed framework,

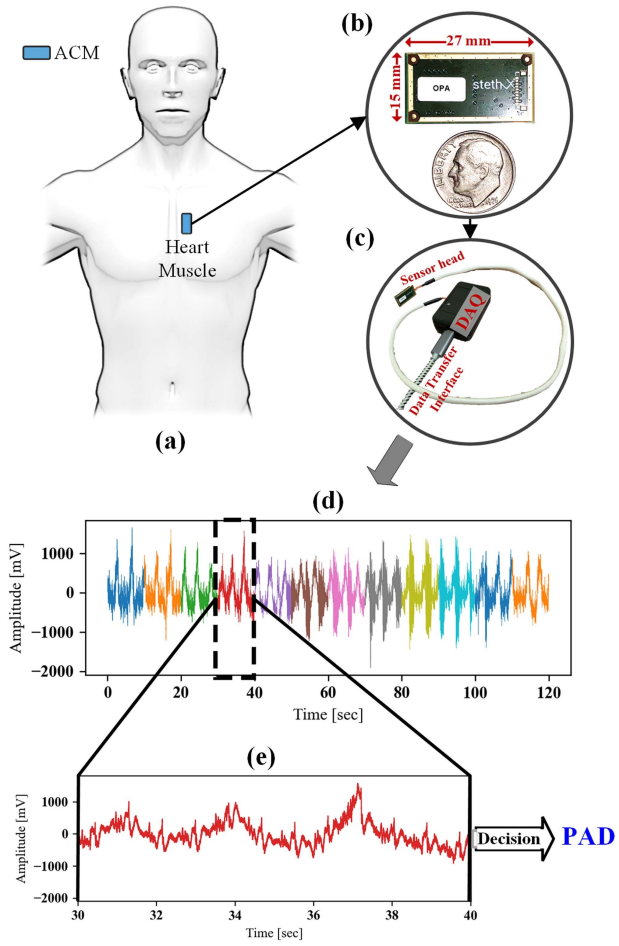

Fig. 1. The proposed PAD detection framework. (a) The sensor ar-rangement on the torso, (b) dimensions of the accelerometer contact microphone, (c) the data acquisition module, (d) the ACM signal seg-mentation, and (e) decision making based on a signal segment.

the objective is to detect peripheral artery disease using short segments of ACM recordings as shown in Fig. 1(d) and (e).

B. Data Preparation

In order to reduce the computational complexity of the pro-cessing chain, ACM recordings were initially down-sampled to a sampling rate of 4 kHz. This practice did not affect the heart sound components which fall below 1 kHz [27]. The ACM sig-nals were then segmented into 10-second overlapping windows with 90% overlap between consecutive windows. The choice of the window length follows the superior results achieved for the 10-second duration compared to other segment lengths in [28]. Depending on the sensor placement and the body mass index (BMI) of the subject, the signal-to-noise ratios (SNRs) of the recorded signals vary within the range of [5, 15] dB. This motivates the need for a denoising algorithm to enhance the quality of heart sounds.

Fig. 2 depicts two examples of ACM raw recordings with SNR values of 10.5 dB and 5.3 dB. The denoising algo-rithm begins with zero-padding the beginning and end of each segment by 256 points (62.5 ms) to prevent the transient effect of

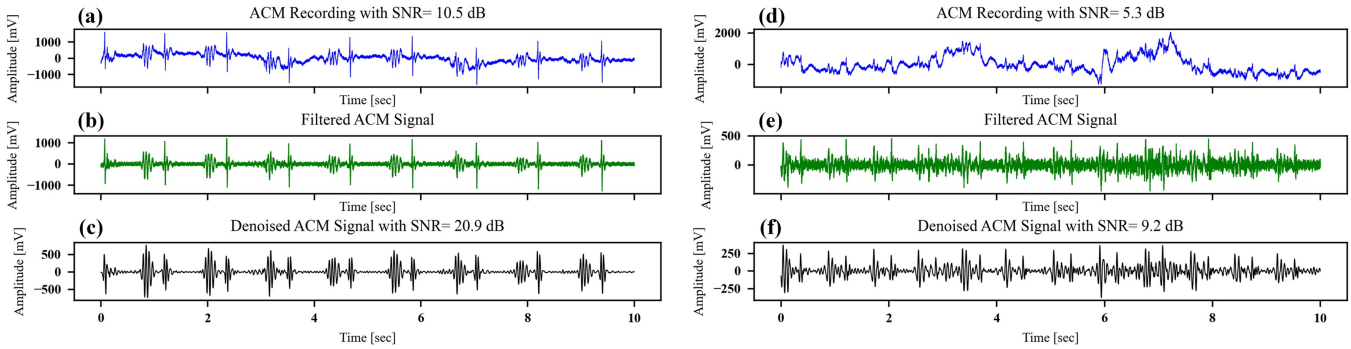

Fig. 2. Signal denoising examples for high-SNR and low-SNR ACM segments. (a) High-SNR ACM raw recording (SNR = 10.5 dB), (b) filtered high-SNR ACM signal segment, (c) denoised high-SNR signal segment generated by EMD, (d) low-SNR ACM raw recording (SNR = 5.3 dB), (e) filtered low-SNR ACM signal segment, and (f) denoised low-SNR signal segment generated by EMD.

filtering conducted in the following processing steps. As shown in Fig. 2(d), heart sound components of the low-SNR phonocardiogram signal are represented by small peaks modulated on the low-frequency waveforms corresponding to respiration. A differentiation filter was applied to the signal segment to amplify high-frequency components. Each segment was then high pass-filtered by a 3rd-order zero-phase Butterworth filter with a cut-off frequency of 10 Hz to remove respiratory components and baseline wander while maintaining heart sound components, which fall above 10 Hz as mentioned in [29]. A cumulative filter was used to reconstruct the heart sound morphology, followed by discarding the padded points at the beginning and end of each segment. Fig. 2(b) and (e) respectively illustrate the filtered ACM segments with SNR 10.5 dB and 5.3 dB. According to Fig. 2(e), the remaining signal is still contaminated with high-energy noise, hindering the full appearance of the first and second heart sounds which are known as S1 and S2, respectively.

Several studies have demonstrated the potentials of empirical mode decomposition (EMD) for phonocardiogram denoising [30], [31]. The core idea behind EMD is to adaptively decompose an oscillatory signal into a series of zero-mean amplitude- and frequency-modulated signals called intrinsic mode functions (IMFs) through an iterative procedure. EMD begins with interpolating between local maxima and local minima to obtain upper and lower envelopes, respectively. These two envelopes are then averaged, and the resulting component is subtracted from the original signal. The resulting signal is called an IMF only if it holds a zero mean, and the number of local extrema and zero crossings are equal or differ by at most 1. The IMF is then subtracted from the original signal to produce the residue signal which plays the role of the original signal for the next iteration. This process repeats until the final residue signal is a constant or monotonic function. Each filtered ACM signal segment (x(n)) undergoes EMD which results in:

$$x(n) = \sum_{i=1}^{L-1} h^i(n) + r^L(n), \tag{4}$$

where $h^i(n)$, $L-1$, and $r^L(n)$ denote the i^{th} IMF, number of IMFs, and the final residue signal respectively. In this work, we propose to denoise the signal segment by selecting the most relevant IMFs. The selection criteria can be defined either in the frequency domain [32] or time domain [30], where the latter has shown promising results for heart sound denoising. Hence, a modified version of the method presented in [30] is designed to incorporate only relevant and smooth IMFs for SNR enhancement. To this end, the actual energy of each IMF is compared with the estimated energy (V_i) proposed by [30], as mentioned below:

$$E_i = \frac{E_1}{\beta}\alpha^{-i};\ i = 2, 3, 4, \ldots, L, \tag{5}$$

where α and β denote the estimation parameters which are 2.01 and 0.719 respectively. E_1 represents the estimated energy for the first IMF which is calculated using the median absolute deviation (MAD) as follows:

$$E_1 = \frac{\left(\text{med}\,h^1 - \text{med}\,h^1\right)^2}{\rho}, \tag{6}$$

with med. and ρ being the median operator and 0.6745 according to [31], respectively. Each IMF whose actual energy exceeds its estimated energy is included for signal reconstruction. In the original denoising method, relevant IMFs are then further denoised using a thresholding-based method to remove suspect noise components. In our signal processing chain however, the desire is to preserve potential patterns corresponding to PAD. To this end, we replace the thresholding technique with a 3rd-order Savitzky-Golay smoothing filter of size 50 ms to produce smooth outputs while maintaining signal patterns [33]. As illustrated in Fig. 2(c) and (f), the denoising algorithm has enhanced the signal quality by 10.4 dB and 3.9 dB for high-SNR and low-SNR signal segments respectively.

C. Feature Extraction

Blood backflow associated with obstructions in PAD patients is expected to manifest abnormal patterns in the energy of ACM recordings and their corresponding dynamics. Inspired by cardiologists who use stethoscopes to inspect the rhythm and intensity of the heart sounds, we extract energy-related features to automatically diagnose PAD. These features are summarized as follows:

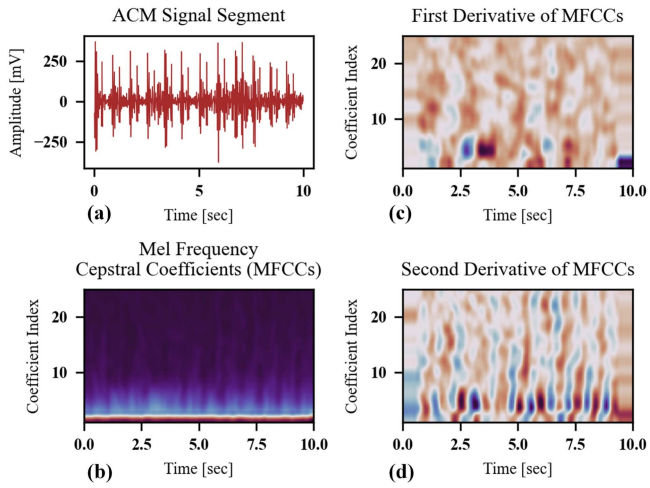

Fig. 3. Heart sound feature extraction. (a) denoised ACM signal segment, (b) Mel frequency cepstral coefficients (MFCCs), (c) the first derivative of MFCCs, and (d) the second derivative of MFCCs.

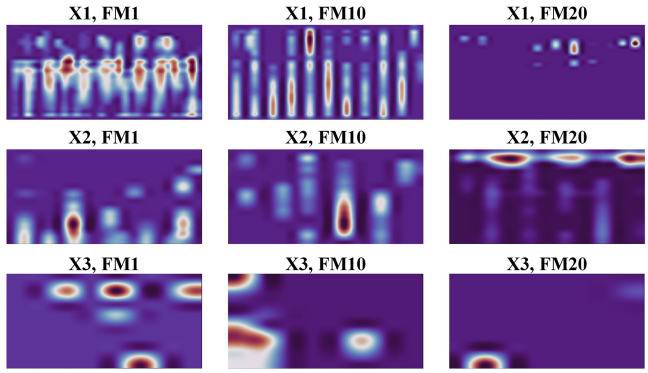

Fig. 4. Examples of high-level-of-abstraction coefficients extracted from MFCCs and its derivatives by ResNet50. Top, middle, and bottom rows correspond to feature maps (FM1, FM10, FM20) from layers 10, 22, and 40 respectively.

1) Mel Frequency Cepstral Coefficients (MFCCs): These coefficients distribute the short-term power spectrum of a signal in accordance with the human auditory system perception in a scale called Mel [34]. Hence, the use of MFCCs allows for replicating auscultation monitoring performed by cardiologists for the detection of abnormal heart sounds. Mel scale uses a filter spaced linearly at frequencies below 1 kHz and has logarithmic spacing above 1 kHz. MFCC computation begins with framing a signal segment into shorter frames of 0.5 seconds (2,048 samples) with 25% overlap (512 samples) between consecutive frames. For each frame, the periodogram is then calculated and filtered through a Mel filter bank with 25 triangular filters. The energy of the signal is summed per filter, thus generating 25 coefficients. Finally, the discrete cosine transform is applied to the coefficients to obtain MFCCs. Hence, for each 10-second signal segment, we would have a 2-D representation of size 25 × 81, representing 81 frames, each involving 25 coefficients. Fig. 3(a) and (b) depict a signal segment and its corresponding MFCCs respectively. In order to model the dynamics of the heart sounds, the first and second derivatives of MFCCs are calculated by subtracting the coefficients of successive time frames once for the first derivative and twice for the second derivative as shown in Fig. 3(c) and (d), respectively. Time differentiation in signal processing acts as a high-pass filter, amplifying the patterns of abrupt changes of energy associated with heart sound abnormalities.

2) High-Level-of-Abstraction Coefficients (HLACs): Convolutional neural networks (CNNs) have shown excellent promise for image-related tasks such as image classification [35], segmentation [36], and denoising [37]. In addition, they are also capable of extracting discriminative features from raw data at different levels of abstraction [38]. We use this capability of CNNs to extract features from MFCCs which can be considered as an input image to a 2-D CNN architecture. To this end, we adopt a pre-trained residual network with 50 layers (ResNet50) [39], and fine-tune the network on our own dataset. This practice, which is called transfer learning, allows us to reuse ResNet50 pre-trained on the ImageNet dataset for the PAD detection task. Leveraging residual layers in ResNet prevents vanishing gradient issues which generally occur in deep networks. As such, ResNet50 is capable of providing features with a variety of abstraction levels. In this work, we use the outputs of layers 10, 22, and 40 of a ResNet50 which have 256, 512, and 1,024 feature maps (FMs), respectively. Fig. 4 depicts a few examples of feature maps 1, 10, and 20, corresponding to layers 10, 22, and 40 which are represented by X1, X2, and X3 respectively. As can be observed from top to bottom in Fig. 4, transfer learning-based HLACs provide both fine and coarse features of the input, helping with decision-making about the presence of PAD. As a result, three feature maps associated with MFCCs and their first and second derivatives, denoted by X, are combined with HLACs to contribute to the accurate detection of PAD.

D. Multi-Stream-Powered Context-Aware Prediction

As mentioned in Section II, the cardiac muscle should pump the blood harder to cope with backflow pressure waves caused by PAD. Abnormal changes in pressure and thus energy level of the ACM recordings are the keys to PAD detection, for which MFCCs and HLACs are derived. The abnormality associated with PAD may appear in any sub-band or time point. This motivates the need for a 2-D context-aware predictive model to characterize inter-dependencies within various sections of a feature map.

Transformers are widely used context-aware models in natural language processing [40], which outperform their ancestors such as long short-term memory (LSTM) networks which were formerly adopted for machine translation tasks and time series modeling. The advantage of transformers over conventional methods is that sequences are modeled as a whole using self-attention modules, whereas LSTM models a sequence on an element-wise basis. In our application however, we cannot use the original structure of the transformer since input features i.e., MFCCs and HLACs, are two-dimensional features. Vision transformer (ViT) is a recently proposed transformer which is customized

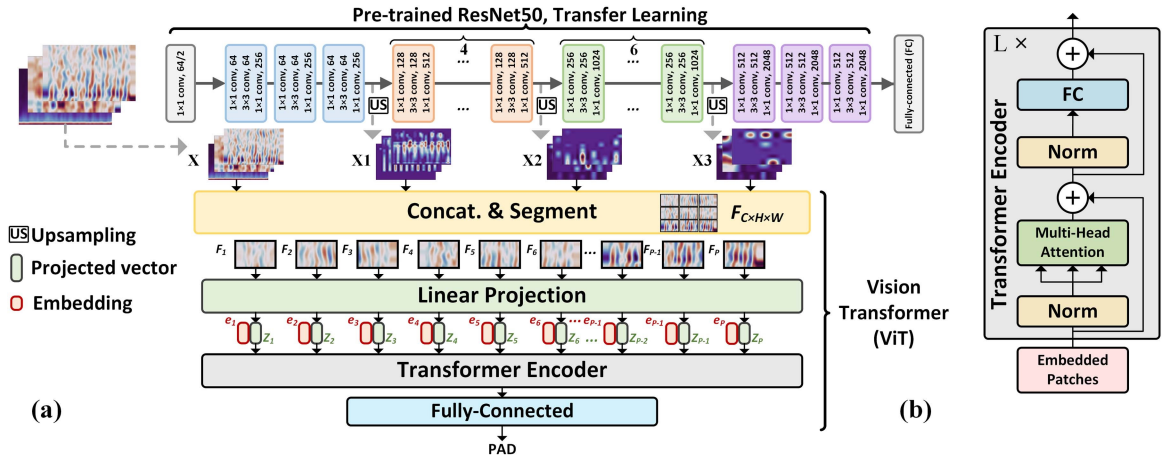

Fig. 5. Multi-stream-powered vision transformer (MSPViT) architecture for PAD detection network. (a) a ResNet50 for HLACs extraction followed by a vision transformer (ViT) for PAD detection, and (b) a transformer encoder which employs a multi-head attention module for modeling inter-patch dependencies.

for image classification [41]. In this work, we develop a network based on ViT which is powered by a multi-stream ResNet50 for PAD detection, as illustrated in Fig. 5.

As shown in Fig. 5(a), a pre-trained ResNet50 on the 1,000-class ImageNet dataset is fine-tuned on the PAD dataset. For this task, the fully-connected (FC) layer at the end of the original ResNet50 with a length of 1,000 is replaced with a fully-connected layer with either two neurons for PAD detection, or three neurons for severity classification. As mentioned in Section III-C2, three MFCCs-related feature maps (X) and the outputs of layers 10 (X1), 22 (X2), and 40 (X3) which are HLACs feature maps, are concatenated to build the input feature maps to the vision transformer network. As mentioned earlier, the dimension of MFCCs is 25 × 81. We resize MFCCs and the corresponding first and second derivatives through a bilinear interpolation to 128 × 128 to have standard-size inputs. This results in X1, X2, and X3 being of various sizes of 32 × 32, 16 × 16, and 8 × 8, respectively. All feature maps X1, X2, and X3 should therefore be reshaped to 128 × 128 using bilinear interpolation, so these features would be concatenable with the resized MFCCs-related features. As a result, the concatenation builds a larger set of feature maps consisting of 1,795 (3 + 256 + 512 + 1,024) feature maps of size 128 × 128.

As shown in Fig. 5(a), concatenated feature maps are denoted by $F \in \mathrm{R}^{C\times H\times W}$ with C, H, and W representing the number of channels (1,795), height (128), and width (128) of the feature maps, respectively. In the vision transformer, F is segmented into P small patches represented by $F_i \in \mathrm{R}^{C\times p\times p}$, $i = 1, 2, 3, \ldots, P$, where p shows the dimensions of the patches. Each $p \times p$ patch is then reshaped to a vector with a length $C \times p \times p$. This practice allows for processing images as 1-D vectors, similar to the processing of time series using a transformer network. In this work, the size of each patch is set to $p = 8$, thus resulting in 256 patches for an input feature size of 128 × 128. A linear projection layer is then applied to each 114,800-long flattened vector to shrink it to a vector of length 64. As a result, P vectors of size 32 ($Z_i \in \mathrm{R}^{1\times 64}$, $i = 1, 2, 3, \ldots, P$) are generated, and summed with trainable embedding vectors which are denoted by $e_i \in \mathrm{R}^{1\times 64}$, $i = 1, 2, 3, \ldots, P$ in Fig. 5(a). The embeddings enable the network to take the position of each patch in the image into account while learning the patterns of healthy and PAD subjects.

To encode the dependencies among the patches of a whole sequence, we employ L = 4 stacked transformer encoders, the architecture of each is illustrated in Fig. 5(b). This encoder consists of two normalization layers, 8 multi-head attention layers, and a fully-connected (FC) as well as two residual layers. Normalization layers accelerate the learning process and enhance the generalizability of the model [42]. After the batch normalization step, a multi-head attention module is employed to quantify the dependencies among the embeddings corresponding to patches. The attention mechanism begins with multiplying each embedding with three learnable matrices $W^Q \in \mathrm{R}^{32\times 64}$, $W^K \in \mathrm{R}^{32\times 64}$, and $W^V \in \mathrm{R}^{32\times 64}$. This generates three vectors called query ($q_i \in \mathrm{R}^{1\times 32}$), key ($k_i \in \mathrm{R}^{1\times 32}$), and value ($v_i \in \mathrm{R}^{1\times 32}$) for each embedding. The query of each embedding is multiplied by all the keys in the sequence to find the relevance between the embedding under test and other elements in the sequence. The score vector which is of size P subsequently undergoes a Softmax function to rescale the scores into the range of [0, 1], and make them sum up to 1. The resulting rescaled scores are then multiplied by their corresponding value vectors to build the attention module output for the embedding under test. Hence, the attention module can be formulated as:

$$U = \text{Softmax}\left(\frac{Q \times K^T}{\sqrt{32}}\right) V, \tag{7}$$

where $Q \in \mathrm{R}^{P\times 32}$, $K \in \mathrm{R}^{P\times 32}$, and $V \in \mathrm{R}^{P\times 32}$ denote the matrices built by stacking query vectors, key vectors, and value vectors respectively. Matrix $U \in \mathrm{R}^{P\times 32}$ shows the output matrix with P rows, each of which represents the output of the attention layer corresponding to a patch (or embedding). In our implementation, we embed 8 attention heads for each attention layer, yielding 8 attention outputs U_i, $i = 1, 2, 3, \ldots, 8$. Per embedding then, a vector of size 8 × 32 results. Every output

is finally converted to a vector of size 256 through a fully-connected layer in the transformer encoder as shown in Fig. 5(b). Another fully connected layer is employed in the PAD detection network in Fig. 5(a) to resize the encoder output to the number of classes, i.e., two for PAD identification and three for PAD severity level classification. The objective of the multi-stream-powered vision transformer (MSPViT) is to minimize the error between predicted values and the corresponding ground truth labels, using mean squared error (MSE) as follows:

$$\ell(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \tag{8}$$

where y and $\hat{y}$ denote one-hot coding vectors for the ground truth and predicted values, respectively.

IV. Experimental Results

This section provides an overview of the training procedure and the metrics used to evaluate the performance of MSPViT. Additionally, the performance of the model on PAD detection and severity classification is discussed in the following sub-sections.

A. Training Procedure

As mentioned in Section III-C, each signal segment is represented by its MFCCs and HLACs which are used as the input to the MSPViT. This network is trained in a supervised manner to learn the subjects' labels assigned by two interventional cardiologists based on echocardiography parameters. In this work, we trained the model for both the detection of PAD (binary classification) and the classification of PAD severity levels into either healthy, mild PAD, or severe PAD (ternary classification).

The training procedure was conducted using the data of 95 subjects amounting to 46,878 signal segments. A 10-fold leave-subject-out cross-validation (10-LSOCV) approach was adopted to assess the generalizability of the model to unseen data. To this end, the dataset was split into train (68 subjects; 70%), validation (18 subjects; 20%), and test (9 subjects; 10%) sets. The model is trained on the train set for 150 epochs, validated on the validation set, and the performance is reported using the test set when the loss function, defined in (8), is minimized. The Adam optimization algorithm [43] was employed with an initial learning rate of 0.001 which was reduced by a factor of 0.98 every ten epochs without performance improvement. The model was implemented on a 24-GB NVIDIA GeForce RTX 3090 Ti FTW3 with a batch size of 64.

B. Evaluation Metrics

Statistical analyses were conducted to evaluate the performance of MSPViT and compare it with state-of-the-art PAD detection methods. In order to compare the ground-truth labels with those predicted by MSPViT, sensitivity (Sen), specificity (Spec), accuracy (Acc), positive predictive value (PPV), and F1 score (F1) are calculated as follows:

$$\text{Sen} = 100 \times \frac{TP}{TP + FN}, \tag{9}$$

TABLE II
10-LSOCV Average (±Standard Deviation) Performance on PAD Detection

Architecture	Sen (%)	Spec (%)	Acc (%)	PPV (%)	F1 (%)
ResNet50 (X)	92.69 (± 4.51)	90.27 (± 11.44)	91.55 (± 4.99)	96.04 (± 4.48)	94.24 (± 3.40)
ViT (X)	99.08 (± 0.71)	93.36 (± 5.48)	97.52 (± 1.87)	97.61 (± 2.66)	98.32 (± 1.36)
MSPViT (X1)	99.05 (± 0.84)	98.96 (± 1.0)	99.05 (± 0.67)	99.68 (± 0.31)	99.36 (± 0.47)
MSPViT (X2)	**99.45 (± 0.47)**	**98.21 (± 2.93)**	**99.07 (± 0.89)**	**99.30 (± 1.25)**	**99.37 (± 0.63)**
MSPViT (X3)	98.92 (± 0.97)	93.47 (± 7.74)	97.20 (± 2.50)	97.31 (± 3.58)	98.07 (± 1.79)
MSPViT (X, X1, X2, X3)	98.66 (± 1.02)	96.85 (± 2.76)	98.04 (± 0.97)	98.68 (± 1.53)	98.65 (± 0.73)

$$\text{Spec} = 100 \times \frac{TN}{TN + FP}, \tag{10}$$

$$\text{Acc} = 100 \times \frac{TP}{TP + FP + FN}, \tag{11}$$

$$\text{PPV} = 100 \times \frac{TP + TN}{TP + FP}, \tag{12}$$

$$\text{F1} = \frac{2 \times \text{PPV} \times \text{Sen}}{\text{PPV} + \text{Sen}}, \tag{13}$$

where TP, TN, FP, and FN denote the true positive (correctly detected as PAD), true negative (correctly detected as healthy), false positive (falsely detected as PAD), and false negative (falsely detected as healthy) cases, respectively. It is to be noted that statistical metrics were calculated for each label in the ternary classification, and each metric was averaged over all three classes. i.e., macro averaging. Sensitivity and specificity respectively indicate the concordance of MSPViT with respect to the reference, whereas the positive predictive value signifies the likelihood that MSPViT can successfully identify PAD. Additionally, the F1 score was adopted to present a harmonic mean of positive predictive value and sensitivity of MSPViT.

C. PAD Detection Performance

Table II summarizes the results of 10-LSOCV for PAD detection (binary classification), where the test dataset at each fold consists of 6 PAD patients and 3 healthy subjects. In this table, MSPViT (X1), MSPViT (X2), and MSPViT (X3) respectively denote vision transformers supplied with the feature maps of layers 10, 22, and 40 of a ResNet50 fine-tuned on the PAD data. The last row signifies the results associated with MSPViT (X, X1, X2, X3) whose input incorporates MFCCs and their derivatives (three feature maps) as well as the feature maps corresponding to layers 10, 22, and 40 of a pre-trained ResNet50, amounting to a total of 1,795 feature maps. Additionally, the performance of a ResNet50-alone and a vision transformer, denoted by ResNet50 (X) and ViT (X) respectively, are listed in the table for the sake of comparison.

TABLE III
10-LSOCV AVERAGE (±STANDARD DEVIATION) PERFORMANCE ON PAD SEVERITY CLASSIFICATION

Architecture	Sen (%)	Spec (%)	Acc (%)	PPV (%)	F1 (%)
ResNet50 (X)	95.27 (± 5.15)	96.89 (± 2.46)	97.25 (± 2.11)	96.07 (± 4.04)	95.58 (± 4.72)
ViT (X)	89.16 (± 6.53)	94.30 (± 3.61)	95.38 (± 3.34)	94.66 (± 2.98)	91.12 (± 5.09)
MSPViT (X1)	96.31 (± 1.59)	95.94 (± 1.58)	96.22 (± 1.53)	93.72 (±2.79)	94.70 (± 1.97)
MSPViT (X2)	97.60 (± 1.12)	96.89 (± 1.36)	96.97 (± 1.32)	93.61 (± 3.34)	95.18 (± 2.06)
MSPViT (X3)	**96.66 (± 4.08)**	**97.34 (± 2.19)**	**97.65 (± 1.67)**	**96.24 (± 2.51)**	**96.29 (± 3.01)**
MSPViT (X, X1, X2, X3)	96.26 (± 1.42)	96.92 (± 1.19)	97.16 (± 1.13)	95.28 (± 2.68)	95.66 (± 1.92)

As shown in the table, MSPViT (X2) with 99.45% (±0.47) offers the highest sensitivity among other architectures for PAD detection, implying the lowest missed rate of PAD cases. ViT (X) and MSPViT (X1) respectively with 99.08% (±0.71) and 99.05% (±0.84) of sensitivity scores are the runner-ups for PAD detection. MSPViT (X1) and MSPViT (X2) identify PAD with specificity scores of 98.96% (±1.0) and 98.21% (±2.93) respectively, which are superior to those of other architectures. In parallel with specificity, MSPViT (X1) with 99.68% (±0.31) and MSPViT (X2) with 99.37% (±0.63) have the highest PPV, thus are most likely to successfully detect PAD. According to the reported F1 scores, vision transformers powered by high-level-of-abstraction features from layers 10 and 22, i.e., MSPViT (X1) and MSPViT (X2), predict the presence of PAD with 99.36% (±0.47) and 99.37% (±0.63) respectively. The highest F1 score, i.e., 99.37%, implies increases of 5.13% and 1.05% compared to the ResNet50-alone and ViT-alone architectures, respectively. MSPViT (X1) and MSPViT (X2) therefore, suggest the best performance for PAD detection.

D. Severity Classification Performance

Statistical analyses on PAD severity classification are presented in Table III, where the test dataset at each fold consists of 3 mild-PAD, 3 severe-PAD, and 3 healthy subjects. Similar to PAD detection, we quantify the performance of MSPViT in comparison to ResNet50 (X) and ViT (X) using sensitivity, specificity, accuracy, PPV, and F1 scores. As summarized in Table III, the highest sensitivity scores were achieved by MSPViT (X2) and MSPViT (X3) with 97.60% (±1.12) and 96.66% (±4.08) respectively. An average specificity of 97.37% (±2.19) ranks the MSPViT (X3) architecture at the top for false stratification of PAD severity. MSPViT (X3) and MSPViT (X, X1, X2, X3) suggest the highest average PPV scores of 96.24% (±2.51) and 95.28% (±2.68), respectively, presenting superior performance compared to other architectures with 96.07% (±4.04), 94.66% (±2.98), 93.72% (±2.79), and 93.61% (±3.34) for ResNet50 (X), ViT (X), MSPViT (X1), and MSPViT (X2) respectively. In terms of F1 score, MSPViT with 96.29% (±3.01) outperforms MSPViT (X, X1, X2, X3), ResNet50 (X), ViT (X), MSPViT (X1), and MSPViT (X2) with 95.66% (±1.92), 95.58% (±4.72), 91.12% (±5.09), 94.70% (±1.97), and 95.18% ±2.06) respectively, which demonstrates the superiority of MSPViT (X3) for the classification of PAD severity levels.

Comparing Tables II and III signifies the importance of HLACs features as well as the functionality of the layers for PAD detection and severity classification. While layer 22 (lower abstraction) provides a discriminative feature space for PAD detection, i.e., binary classification, severity level classification could benefit from higher abstraction level features in layer 44. This behavior can be related to the ternary classification defined for PAD severity stratification which requires more detailed information in comparison to PAD detection.

E. Receiver Operating Characteristic (ROC) Analysis

The discrimination power of MSPViT is also evaluated using the area under the curve (AUC) of the receiver operating characteristic (ROC). In this work, ROC is defined as the true positive rate versus the false positive rate for each class as illustrated in Fig. 6. The predicted labels of the test data from all folds were appended, and the ROC curve was obtained using all the test data. Fig. 6(a) indicates the ROC curve associated with PAD detection (binary classification), whereas Fig. 6(b), (c), and (d) correspond to the ROC curves of healthy, mild-PAD, and severe-PAD classes respectively. The highest AUC for PAD detection was achieved by MSPViT (X1) with 0.997, followed by MSPViT (X3) and MSPViT (X2) with AUC of 0.994 and 0.990 respectively. ResNet50 (X) offers the lowest AUC of 0.923 which still implies excellent prediction stability. The performance obtained in Fig. 6(a) confirms the validity of the assumption of backflow pressure resulting from arterial obstruction associated with PAD in the arterial tree, mentioned in Section II.

In order to present ROC characteristics in PAD severity classification, we compare each class with others in a one-versus-rest manner. For this purpose, the target class was labeled with one, and the other two classes were assigned zero. As such, three ROC curves associated with healthy, mild-PAD, and severe-PAD classes were achieved as shown in Fig. 6(b), (c), and (d) with AUCs within the ranges of [0.926, 0.992], [0.941, 0.987], and [0.950, 0.978], respectively. Among the MSPViT architectures, the highest AUC scores for healthy, mild-PAD, and severe-PAD cases were achieved by MSPViT (X2) with AUC = 0.989, MSPViT (X2) with AUC = 0.969, MSPViT (X, X1, X2, X3) with AUC = 0.973 respectively, although MSPViT (X3) with $AUC_{healthy}$ = 0.980, AUC_{mild} = 0.941, and AUC_{severe} = 0.950 suggests slight differences compared to their relative superior models. Hence, MSPViT (X3) predicts the severity level with an average AUC of 0.957.

V. DISCUSSION

In the previous section, we comprehensively investigated the feasibility of the detection of PAD and classification of the severity levels through attention-based modules supplied by MFCCs and HLACs. In this section, the clusters corresponding to different classes in the output of the MSPViT are measured to

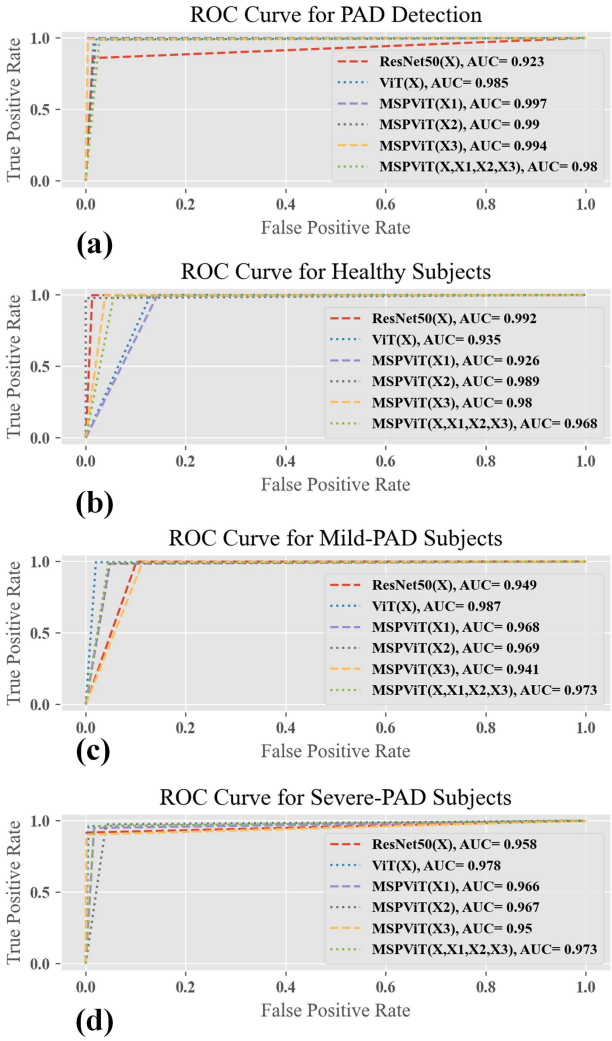

Fig. 6. Receiver operating characteristic (ROC) curves for (a) PAD binary classification, (b) healthy class in ternary classification, (c) mild-PAD class in ternary classification, and (d) severe-PAD class in ternary classification.

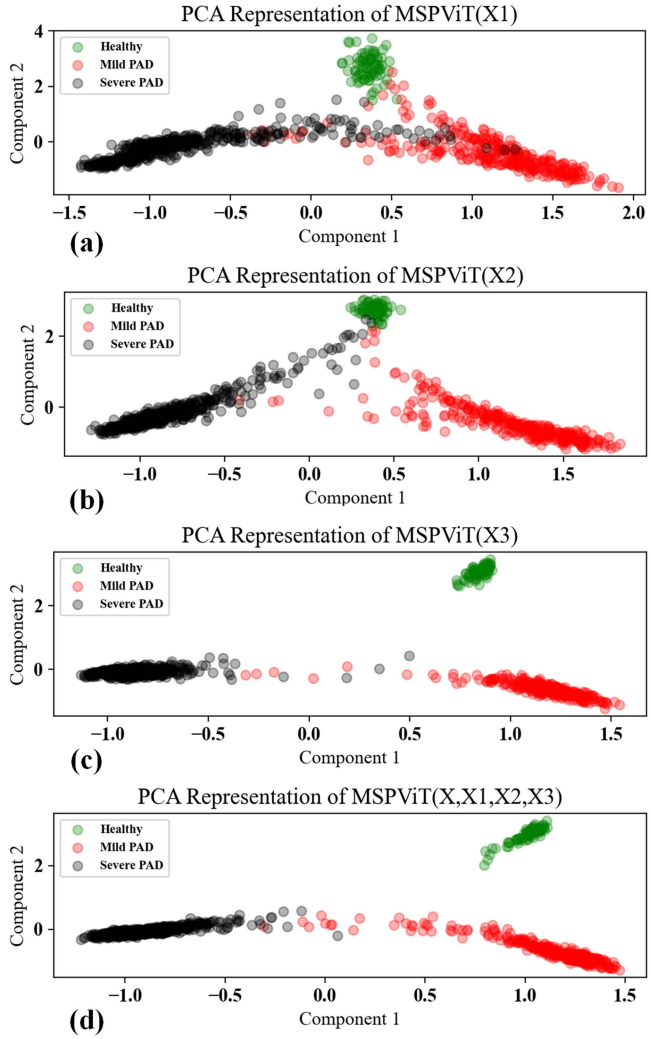

Fig. 7. Principal component analysis (PCA) applied to the outputs of the transformer encoder corresponding to (a) MSPViT (X1), (b) MSPViT (X2), (c) MSPViT (X3), and (d) MSPViT (X, X1, X2, X3).

analyze the separability of classes in a high-dimensional feature space. Additionally, our framework for PAD detection and classification is compared with state-of-the-art methods. Finally, the limitations of the methods are comprehensively discussed.

A. High-Dimensional Separability

A well-trained deep neural network classifier is expected to discriminate test samples of different classes in the high-dimensional space presented by the last layer of the network. We use the output of the transformer encoder layer to inspect separability among PAD classes. Principal component analysis (PCA) was utilized for mapping 32-dimensional feature vectors into a 2-dimensional space as shown in Fig. 7 to visually inspect the separability for fold 1 of test data. As observed in Fig. 7(a), (b), (c), and (d) which respectively correspond to architectures MSPViT (X1), MSPViT (X2), MSPViT (X3), and MSPViT (X, X1, X2, X3), excellent separability is offered among healthy, mild-PAD, and severe-PAD classes. Slight overlaps between classes are due to PCA discarding non-principal components which potentially provide discriminating capabilities. As such, we employ the whole feature vectors which are of length 32, and quantify the clusters formed by each class in the 32-dimensional space. To this end, we adopt the silhouette score for each test data point which is defined as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{14}$$

where $b(i)$ and $a(i)$ denote the distance between a sample and the nearest cluster that the sample does not belong to and the distance between a sample and the cluster that the sample is a part of, respectively. Silhouette score ranges within $[-1, +1]$ with $+1$ suggesting a perfect match between the sample and the assigned label, i.e., healthy, mild-PAD, and severe-PAD. In this work, we present the silhouette score by averaging over the test data points corresponding to 10-LSOCV. Tables IV and V summarize silhouette scores for folds of PAD detection and

TABLE IV
SILHOUETTE SCORES OF OBTAINED CLUSTERS FOR PAD DETECTION NETWORK IN 10-LSOCV

Fold	MSPViT (X1)	MSPViT (X2)	MSPViT (X3)	MSPViT (X, X1, X2, X3)
1	0.75	0.82	0.76	0.75
2	0.40	0.55	0.49	0.42
3	0.76	0.83	0.82	0.79
4	0.77	0.83	0.79	0.80
5	0.48	0.57	0.63	0.62
6	0.78	0.77	0.89	0.85
7	0.45	0.60	0.41	0.42
8	0.39	0.41	0.45	0.37
9	0.41	0.44	0.38	0.44
10	0.75	0.77	0.84	0.75
Average (± Std.)	**0.59 (± 0.18)**	**0.66 (± 0.16)**	**0.65 (± 0.20)**	**0.65 (± 0.18)**

TABLE V
SILHOUETTE SCORES OF OBTAINED CLUSTERS FOR SEVERITY CLASSIFICATION NETWORK IN 10-LSOCV

Fold	MSPViT (X1)	MSPViT (X2)	MSPViT (X3)	MSPViT (X, X1, X2, X3)
1	0.63	0.66	0.72	0.76
2	0.66	0.74	0.77	0.73
3	0.56	0.66	0.87	0.83
4	0.59	0.67	0.84	0.75
5	0.71	0.70	0.81	0.71
6	0.70	0.65	0.86	0.75
7	0.68	0.70	0.81	0.73
8	0.63	0.76	0.83	0.69
9	0.67	0.77	0.81	0.75
10	0.59	0.71	0.77	0.78
Average (± Std.)	**0.70 (± 0.04)**	**0.70 (± 0.05)**	**0.81 (± 0.03)**	**0.75 (± 0.04)**

severity classification, respectively. Table IV suggests the superiority of MSPViT (X2) with a silhouette score of 0.66 (±0.16) over MSPViT (X1), MSPViT (X3), and MSPViT (X, X1, X2, X3) with AUCs of 0.59 (±0.18), 0.65 (±0.20), and 0.65 (±0.18) respectively, as also concluded in Table II. As mentioned in Table V, MSPViT (X3) with a silhouette score of 0.81 (±0.03) outperforms MSPViT (X1), MSPViT (X2), and MSPViT (X, X1, X2, X3) with 0.70 (±0.04), 0.70 (±0.05), and 0.75 (±0.04), respectively, in classifying the severity levels. These results are consistent with our findings in Table III where an F1 score of 96.29% (±3.01) in severity classification is reported for MSPViT (X3). Comparing Tables IV and V implies higher dissimilarities among clusters in severity classification than PAD detection. This finding is aligned with the example shown in Fig. 7, where mild and severe PAD classes fall apart in the two-dimensional space. The distance between mild and severe PAD classes in the 32-dimensional feature space, while being considered in the same cluster, leads to a lower silhouette score. Additionally, this observation confirms the difference between the nature of mild and severe cases in a high-dimensional space.

B. Comparative Analysis

Table VI compares MSPViT with state-of-the-art methods for PAD detection. These methods leverage either blood pressure patterns acquired from an arm cuff [13], [14], or PPG sensors placed on toes [16], [17], [19] to recognize PAD patients. As discussed in Section I, state-of-the-art methods employ deep learning (DL) techniques and statistical analysis to process PPG and oscillometry recordings for the detection of PAD.

Our proposed framework for PAD detection, ACM + MSPViT (X2), outperforms other methodologies in terms of sensitivity (99.4% vs. 90.0%, 90.0%, 93.0%, 92.0%, and 82.4% respectively offered by [13], [14], [16], [17], and [19]), implying the lowest missed PAD rate. The specificity of the proposed method, i.e., 98.2%, exceeds the oscillometry method in [13] by a margin of 0.8%, suggesting a lower false detection rate. MSPViT (X2) offers an accuracy of 99.1% which is considerably greater than 94.8%, 91.4%, 91.0%, and 86.9% presented by [13], [14], [16], and [19], respectively. Comparing the F1 score of the proposed method with that reported in [13] (99.3% vs. 92.3%), an improvement of 7.0% is concluded as a result of using ACM proximal recordings with MSPViT (X2). An area under the curve (AUC) of 0.99 by ACM + MSPViT (X2) outperforms the oscillometry method in [13] and the PPG-based method in [17] with AUC scores of 0.94 and 0.79 respectively. Additionally, our proposed methodology is capable of determining the severity level of PAD with an F1 score of 96.3% and an average AUC score of 0.96, whereas conventional methods are unable to evaluate the severity level.

C. Limitations and Future Works

It was demonstrated that patterns representing the presence of PAD and its corresponding severity level could be detected using proximal accelerometer contact microphone recordings. While our method offers excellent performance, it could benefit from addressing the limitations mentioned in the following paragraphs.

As mentioned in Section III-A, the ACM sensor should have high precision and sensitivity to record minute vibrational patterns corresponding to PAD. This is however accompanied by recording undesired components such as respiration and vocal cord vibrations due to speaking. Fig. 8 represents an example of a signal distorted by vocal cord vibrations. This distortion affects feature extraction and hence the decision-making process. This problem can be addressed through the use of blind source separation (BSS) techniques to discriminate cardiac components from vocal vibrations. The BSS method could be either based on a single-channel platform similar to our previous work in [44], or an array of sensors used for the independent component analysis (ICA) technique [45].

The purpose of this study is to make a decision on the presence of PAD as well as classify its severity levels. However, the proposed framework is unable to locate the obstruction region, which if possible, can be followed by medical interventions. In

TABLE VI
COMPARATIVE ANALYSIS OF PAD DETECTION AND SEVERITY CLASSIFICATION BASED ON ACCELEROMETER CONTACT MICROPHONE DATA TRAINED ON MSPVIT ARCHITECTURES VERSUS STATE-OF-THE-ART

Methodology	Data Type	Task	Number of PAD Patients	Performance				
				Sen (%)	Spec (%)	Acc (%)	F1 score (%)	AUC
Oscillometry + DL [13]	Arterial pressure wave	PAD detection	14	90.0	97.4	94.8	92.3	0.94
Oscillometry + DL [14]	Arterial pressure wave	PAD detection	14	90.0	92.1	91.4	-	-
PPG + Statistical Analysis [16]	Bilateral toe PPG	PAD detection	48	93.0	89.0	91.0	-	-
PPG + Statistical Analysis [17]	Toe PPG	PAD detection	30	92.0	45.0	-	-	0.79
PPG + DL [19]	Toe PPG	PAD detection	80	82.4	89.0	86.9	-	-
ACM + MSPViT (X2)	**Proximal ACM signals**	**PAD detection**	**74**	**99.4**	**98.2**	**99.1**	**99.3**	**0.99**
ACM + MSPViT (X3)	**Proximal ACM signals**	**PAD severity classification**	**74**	**96.7**	**97.3**	**97.6**	**96.3**	**0.96**

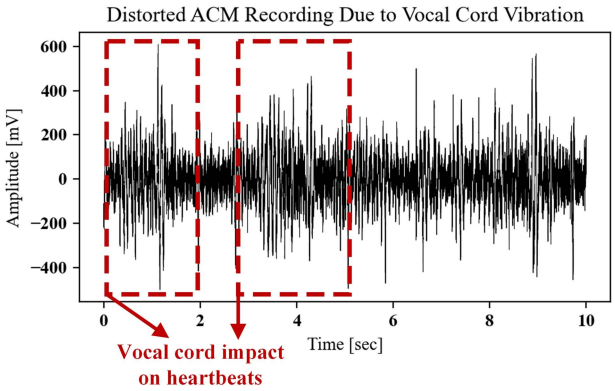

Fig. 8. An example of an ACM recording distorted by vocal cord vibrations during speaking.

our future studies, we will investigate the possibility of determining the obstruction regions using multiple-ACM configurations distributed on the limbs. The signal pattern changes among various arterial sites may provide useful information about the obstruction region(s).

VI. CONCLUSION

This study presents a novel framework based on the proximal recordings of heart vibrations using a high-precision accelerometer contact microphone (ACM) for the detection of peripheral artery disease (PAD) and the classification of its severity level. A deep neural network coined multi-stream-powered vision transformer (MSPViT) is introduced to differentiate PAD patients from healthy subjects using blood pressure backflow patterns in ACM recordings corresponding to arterial obstructions. The proposed architecture leverages Mel frequency cepstral coefficients (MFCCs) and high-level-of-abstraction coefficients (HLACs) generated by a pre-trained ResNet50 network for PAD detection and severity classification. The performance of the proposed framework is evaluated on the data of 74 PAD patients and 21 healthy subjects (a total of 95 subjects) for PAD detection as well as severity classification. Sensitivity, specificity, accuracy, positive predictive value, and F1 scores of 99.45% (±0.47), 98.21% (±2.93), 99.07% (±0.89), 99.30% (±1.25), and 99.37% (±0.63) are respectively reported for PAD detection using MSPViT (X2) whose input is provided by layer 22 of the fine-tuned ResNet50 network. An AUC of 0.99 also confirms the discrimination power of MSPViT (X2). Additionally, the network is assessed for the classification of severity levels where average sensitivity, specificity, accuracy, positive predictive value, and F1 scores of 96.66% (±4.08), 97.34% (±2.19), 97.65% (±1.67), 96.24% (±2.51), and 96.29% (±3.01) are achieved. As such, for the first time in this study, the functionality of proximal recordings of heart sounds for PAD detection is proven.

ACKNOWLEDGMENT

Farrokh Ayazi is an inventor of the ACM technology being used in this study and involved with efforts to explore its commercialization. The terms of this arrangement have been reviewed and approved by Georgia Tech in accordance with its conflict-of-interest policies.

REFERENCES

[1] S. S. Virani et al., "Heart disease and stroke statistics–2021 update: A report from the American Heart Association," *Circulation*, vol. 143, no. 8, pp. e254–e743, 2021.
[2] M. H. Criqui and V. Aboyans, "Epidemiology of peripheral artery disease," *Circulation Res.*, vol. 116, no. 9, pp. 1509–1526, 2015.
[3] M. M. McDermott, L. Fried, E. Simonsick, S. Ling, and J. M. Guralnik, "Asymptomatic peripheral arterial disease is independently associated with impaired lower extremity functioning: The women's health and aging study," *Circulation*, vol. 101, no. 9, pp. 1007–1012, 2000.
[4] E. Selvin and T. P. Erlinger, "Prevalence of and risk factors for peripheral arterial disease in the United States: Results from the national health and nutrition examination survey,1999–2000," *Circulation*, vol. 110, no. 6, pp. 738–743, 2004.
[5] I. J. Kullo and T. W. Rooke, "Peripheral artery disease," *New England J. Med.*, vol. 374, no. 9, pp. 861–871, 2016.
[6] S. A. Carter, "Indirect systolic pressures and pulse waves in arterial occlusive disease of the lower extremities," *Circulation*, vol. 37, no. 4, pp. 624–637, 1968.

[7] R.-A. Marius, L. Iliuta, S. M. Guberna, and C. Sinescu, "The role of ankle-brachial index for predicting peripheral arterial disease," *Maedica*, vol. 9, no. 3, 2014, Art. no. 295.
[8] J. G. Lijmer, M. G. Hunink, J. J. van den Dungen, J. Loonstra, and A. J. Smit, "ROC analysis of noninvasive tests for peripheral arterial disease," *Ultrasound Med. Biol.*, vol. 22, no. 4, pp. 391–398, 1996.
[9] W. Grossman, "Cardiac catheterization and angiography," 3rd ed., U.S., 1986.
[10] T. Laswed et al., "Assessment of occlusive arterial disease of abdominal aorta and lower extremities arteries: Value of multidetector CT angiography using an adaptive acquisition method," *Eur. Radiol.*, vol. 18, no. 2, pp. 263–272, 2008.
[11] R. Schernthaner et al., "Multidetector ct angiography in the assessment of peripheral arterial occlusive disease: Accuracy in detecting the severity, number, and length of stenoses," *Eur. Radiol.*, vol. 18, no. 4, pp. 665–671, 2008.
[12] D. J. Gradinscak, N. Young, Y. Jones, D. O'Neil, and D. Sindhusake, "Risks of outpatient angiography and interventional procedures: A prospective study," *Amer. J. Roentgenol.*, vol. 183, no. 2, pp. 377–381, 2004.
[13] N. Forghani, K. Maghooli, N. J. Dabanloo, A. V. Farahani, and M. Forouzanfar, "DeepPAD: Detection of peripheral arterial disease using deep learning," *IEEE Sensors J.*, vol. 22, no. 16, pp. 16254–16262, Aug. 2022.
[14] N. Forghani, K. Maghooli, N. J. Dabanloo, A. V. Farahani, and M. Forouzanfar, "Intelligent oscillometric system for automatic detection of peripheral arterial disease," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 3209–3218, Aug. 2021.
[15] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiol. Meas.*, vol. 28, no. 3, 2007, Art. no. R1.
[16] J. Allen, K. Overbeck, A. F. Nath, A. Murray, and G. Stansby, "A prospective comparison of bilateral photoplethysmography versus the ankle-brachial pressure index for detecting and quantifying lower limb peripheral arterial disease," *J. Vasc. Surg.*, vol. 47, no. 4, pp. 794–802, 2008.
[17] M. Peltokangas et al., "Parameters extracted from arterial pulse waves as markers of atherosclerotic changes: Performance and repeatability," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 3, pp. 750–757, May 2018.
[18] M. Bentham, G. Stansby, and J. Allen, "Innovative multi-site photoplethysmography analysis for quantifying pulse amplitude and timing variability characteristics in peripheral arterial disease," *Diseases*, vol. 6, no. 3, 2018, Art. no. 81.
[19] J. Allen, H. Liu, S. Iqbal, D. Zheng, and G. Stansby, "Deep learning-based photoplethysmography classification for peripheral arterial disease detection: A proof-of-concept study," *Physiol. Meas.*, vol. 42, no. 5, 2021, Art. no. 054002.
[20] J.-X. Wu, C.-M. Li, Y.-R. Ho, M.-J. Wu, P.-T. Huang, and C.-H. Lin, "Bilateral photoplethysmography analysis for peripheral arterial stenosis screening with a fractional-order integrator and info-gap decision-making," *IEEE Sensors J.*, vol. 16, no. 8, pp. 2691–2700, Apr. 2016.
[21] Q. Yousef, M. Reaz, and M. A. M. Ali, "The analysis of PPG morphology: Investigating the effects of aging on arterial compliance," *Meas. Sci. Rev.*, vol. 12, no. 6, 2012, Art. no. 266.
[22] U. Rubins, A. Grabovskis, J. Grube, and I. Kukulis, "Photoplethysmography analysis of artery properties in patients with cardiovascular diseases," in *Proc. 14th Nordic-Baltic Conf. Biomed. Eng. Med. Phys.*, 2008, pp. 319–322.
[23] W. He, H. Xiao, and X. Liu, "Numerical simulation of human systemic arterial hemodynamics based on a transmission line model and recursive algorithm," *J. Mech. Med. Biol.*, vol. 12, no. 01, 2012, Art. no. 1250020.
[24] R. Drake, A. W. Vogl, and A. W. Mitchell, *Gray's Anatomy for Students E-Book*. Amsterdam, Netherlands: Elsevier, 2009.
[25] P. Gupta, H. Wen, L. Di Francesco, and F. Ayazi, "Detection of pathological mechano-acoustic signatures using precision accelerometer contact microphones in patients with pulmonary disorders," *Sci. Rep.*, vol. 11, no. 1, pp. 1–12, 2021.
[26] P. Gupta, M. J. Moghimi, Y. Jeong, D. Gupta, O. T. Inan, and F. Ayazi, "Precision wearable accelerometer contact microphones for longitudinal monitoring of mechano-acoustic cardiopulmonary signals," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–8, 2020.
[27] N. J. Rennert, R. Morris, and C. C. Barrere, "How to cope with scopes: Stethoscope selection and use with hearing aids and Cls," *Hear. Rev.*, vol. 11, no. 2, pp. 34–41, 2004.
[28] A. Shokouhmand, N. D. Aranoff, E. Driggin, P. Green, and N. Tavassolian, "Efficient detection of aortic stenosis using morphological characteristics of cardiomechanical signals and heart rate variability parameters," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, 2021.
[29] A. K. Dwivedi, S. A. Imtiaz, and E. Rodriguez-Villegas, "Algorithms for automatic analysis and classification of heart sounds–a systematic review," *IEEE Access*, vol. 7, pp. 8316–8345, 2018.
[30] P. Flandrin, P. Goncalves, and G. Rilling, "Detrending and denoising with empirical mode decompositions," in *Proc. 12th Eur. Signal Process. Conf.*, 2004, pp. 1581–1584.
[31] A. H. Salman, N. Ahmadi, R. Mengko, A. Z. Langi, and T. L. Mengko, "Performance comparison of denoising methods for heart sound signal," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst.*, 2015, pp. 435–440.
[32] A. Shokouhmand, C. Antoine, B. K. Young, and N. Tavassolian, "Multimodal framework for fetal heart rate estimation: Fusion of Low-SNR ECG and inertial sensors," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2021, pp. 7166–7169.
[33] A. Savitzky and M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.*, vol. 36, no. 8, pp. 1627–1639, 1964.
[34] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Hoboken, NJ, USA: Prentice-Hall, 1993.
[35] M. Hussain, J. J. Bird, and D. R. Faria, "A study on CNN transfer learning for image classification," in *Proc. U.K. Workshop Comput. Intell.*, 2018, pp. 191–202.
[36] F. Sultana, A. Sufian, and P. Dutta, "Evolution of image segmentation using deep convolutional neural network: A survey," *Knowl.-Based Syst.*, vol. 201, 2020, Art. no. 106062.
[37] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 341–349.
[38] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *Proc. Int. Conf. Artif. Neural Netw.*, 2018, pp. 270–279.
[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[40] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[41] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2020.
[42] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?," in *Proc. 32nd Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2488–2498.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *ICLR (Poster)*, 2015.
[44] A. Shokouhmand and N. Tavassolian, "Fetal electrocardiogram extraction using dual-path source separation of single-channel non-invasive abdominal recordings," *IEEE Trans. Biomed. Eng.*, early access, 2022, doi: 10.1109/TBME.2022.3189617.
[45] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Cambridge, MA, USA: Academic press, 2010.