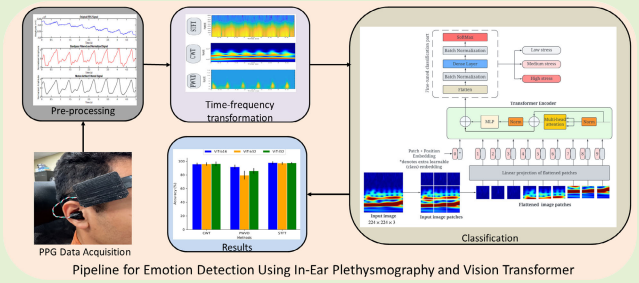
Detection and Classification of Mental Stress Using In-Ear Plethysmography and a Vision Transformer

Hika Barki, Lionel Nkenyereye, and Wan-Young Chung, *Senior Member, IEEE*

***Abstract*—This study addresses the critical need for effective mental stress monitoring, linked to severe health issues like depression and heart disease. We introduce a robust method using in-ear photoplethysmogram (PPG) signals for detecting and classifying stress levels. The objective of this study is to develop a precise stress monitoring technique using advanced signal processing and deep learning. Raw PPG data were collected from 15 subjects undergoing stress-inducing activities in a controlled setting. The data underwent preprocessing and were transformed into image-like time-frequency representations. We employed vision transformer (ViT) models for classification, which were fine-tuned and compared against other state-of-the-art deep learning models. The ViT classifier significantly outperformed existing models, achieving an average accuracy of 97.78% and an *F*1-score of 97.79%. While the dataset is relatively small, these results suggest a promising direction for stress monitoring by illustrating the potential of combining in-ear PPG signals with ViT models. The study indicates the efficacy of this novel approach for accurate mental stress diagnosis, which could have significant implications for mental health applications. Future work will focus on validating these findings with a larger sample size and exploring the integration of this technology into wearable devices for real-world stress monitoring.**



***Index Terms*—Classification, filtering, mental stress, photoplethysmogram (PPG), time-frequency analysis, vision transformer (ViT).**

I. Introduction

MENTAL stress is a prevalent condition in today's fast-paced and demanding world, and it has far-reaching implications for both individuals and society. Mental stress is a type of stress that involves the emotional, psychological, and physiological strain experienced when the demands placed on an individual exceed their ability to cope [1]. Individuals suffering from mental stress frequently experience various symptoms, e.g., anxiety, depression, irritability, and difficulty concentrating, which can significantly impact their overall well-being [2]. In addition, chronic mental stress has been linked to a wide range of physical health consequences, including an increased risk of cardiovascular diseases [3], compromised immune function [4], and higher susceptibility to various illnesses [5]. Beyond the individual level, mental stress can strain interpersonal relationships, which can lead to conflict and a reduced sense of social engagement [6]. It also carries substantial economic implications, e.g., productivity losses, increased healthcare utilization, and the costs associated with stress-related disorders [7]. Thus, understanding the impact of mental stress is critical in terms of developing effective strategies to support individuals and mitigate the broader societal effects of this widespread issue.

Various methods have been developed to detect mental stress in order to identify and assess individuals experiencing heightened psychological strain. For example, the most prevalent method for assessing mental well-being involves the use of questionnaires and consultations with trained professionals [8], [9]. These approaches provide valuable insights into an individual's stress perception, although they are limited by the potential biases and subjectivity inherent in self-reporting techniques. Another approach involves using biomarkers, e.g.,

Received 15 November 2024; accepted 1 December 2024. Date of publication 12 December 2024; date of current version 14 January 2025. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) under Grant 2019R1A2C1089139. The associate editor coordinating the review of this article and approving it for publication was Prof. Liang-Bi Chen. *(Corresponding author: Wan-Young Chung.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Pukyong National University Institutional Review Board under Application No. 1041386-202305-HR-42-02.

Hika Barki and Lionel Nkenyereye are with the Department of AI Convergence, Pukyong National University, Busan 48513, Republic of Korea (e-mail: daljuhika@pukyong.ac.kr; lionelnk82@gmail.com).

Wan-Young Chung is with the Department of Electronic Engineering, Pukyong National University, Busan 48513, Republic of Korea (e-mail: wychung@pknu.ac.kr).

Digital Object Identifier 10.1109/JSEN.2024.3512595

salivary alpha-amylase [10] and cortisol [11]. These kinds of measures offer objective physiological markers; however, specialized equipment and expertise are typically required for data collection and analysis. Consequently, researchers have been diligently developing compact, portable, and accurate technologies for the detection and continuous monitoring of mental stress.

The demand for real-time human health monitoring wearable devices has surged, facilitating mental stress detection through physiological signal measurements. Recent studies suggest that analyzing photoplethysmogram (PPG) signals, a method used for decades in clinical assessments and now available in consumer products like smartwatches, can indicate stress levels [12], [13], [14]. The noninvasive PPG technology measures blood volume changes, providing insights into cardiovascular functions, including heart rate, blood circulation, and autonomic nervous system activity, which are all indicative of stress. Further research connects stress to vascular health, suggesting that PPG signal alterations, reflecting changes in blood perfusion and endothelial function, could form a comprehensive stress index. Such advancements hint at the potential for everyday devices to monitor stress continuously, leveraging established and emerging cardiovascular metrics.

In the domain of mental stress detection through PPG signals, recent research has notably enhanced our understanding and detection capabilities. Notably, a study [15] demonstrated the effectiveness of ultra-short-term PPG recordings for stress detection. The study achieved a notable accuracy rate of 94.33% employing Support Vector Machine classifiers. This underscores the effectiveness of pulse rate variability (PRV) as a practical alternative to heart rate variability (HRV) in assessing stress. Additionally, another study [14] introduced a novel, AI-driven approach for rapid stress assessment using deep learning, employing time-domain and frequency-domain parameters in a GoogLeNet-based model, showcasing the potential of advanced AI methods in stress prediction from PPG signals. Furthermore, research indicated by [16] developed the Orchestrating Multiple Denoising and Peak-Detecting method, combining denoising and peak-detecting techniques to improve PPG signal feature extraction accuracy, achieving an impressive accuracy rate of 96.50% and an $F1$ score of 93.36%. These advancements underscore the growing effectiveness of sophisticated signal processing and machine-learning techniques in enhancing the precision of stress detection through wearable devices.

Advancements in mental stress detection continue, highlighted by studies [17], [18], [19], which each contribute significantly to the field. Study [17] conducted a comparative analysis between PPG and ECG for monitoring HRV, finding moderate to good consistency, emphasizing the need for careful PPG calibration against established ECG benchmarks. In another development, study [18] applied advanced machine learning, using deep neural networks to achieve a 91% classification accuracy in stress detection from PPG signals, demonstrating the potential of sophisticated algorithms. Furthermore, study [19] addressed mental stress in sedentary professions, examining PRV's effectiveness from PPG in evaluating stress among IT professionals across different cognitive states. This research highlighted 18 key features and improved classification accuracy from 85.1% to 91% by shifting from multilayer perceptrons to deep neural networks. Additionally, our previous work [20] introduced stress detection using an ear-in PPG biosensor, applying a convolutional neural network to achieve over 90% accuracy in identifying stress states, showcasing the shift toward real-time, AI-enhanced methods within wearable technology contexts. These contributions mark a significant evolution from elementary to complex, AI-infused PPG signal applications for mental stress detection.

The mentioned studies collectively underscore the significant progress in mental stress detection using PPG signals, highlighting a shift from basic accuracy in initial models to the application of sophisticated machine-learning techniques for improved real-time monitoring in wearable technologies. Yet, there are limitations in current approaches. Traditional cardiovascular-based stress detection, like HRV analysis, is hindered by sensitivity to motion artifacts, which impacts reliability outside controlled environments. Additionally, HRV's dependence on extended monitoring periods challenges its viability for real-time applications. Variability in HRV among individuals, reflecting diverse stress responses, further complicates consistent detection accuracy, as seen in varied HRV responses among different populations, including those with anxiety disorders compared to healthy controls [21]. Also, prevailing stress detection methodologies largely rely on traditional machine learning, necessitating manual feature extraction from physiological signals—a process that might overlook critical signal aspects. These issues underscore the ongoing need for innovative, effective, and adaptable techniques in mental stress detection.

Therefore, we proposed an efficient way of mental stress detection that leverages the unique benefits of in-ear PPG sensing. The in-ear placement has several advantages, such as reduced motion artifacts and a closer connection to cerebral blood flow, potentially improving signal accuracy and sensitivity to stress-induced bodily changes. This method aligns with the trend toward developing more discreet and user-friendly wearable devices for health monitoring. Moreover, our study introduces the application of vision transformer (ViT) models, cutting-edge deep learning techniques initially developed for image data, adapted here for analyzing time-series PPG signals obtained from in-ear sensors. This novel use of ViTs is expected to provide a robust and accurate analysis of stress levels, capturing subtle variations through in-ear PPG. It marks a significant advancement from traditional machine and deep learning methods, offering a more nuanced interpretation of stress markers. Additionally, the combination of in-ear PPG sensing with ViT models is designed to mitigate some existing limitations, such as the need for prolonged signal recordings and handling of noisy data, enabling efficient, real-time stress monitoring suitable for immediate health insights in everyday settings.

The primary contributions of our study are summarized as follows.

1) Integration of a custom PPG sensor with an IMU sensor within an earbud, enhancing mental stress detection capabilities from in-ear measurements.

2) In-ear PPG data collection and examination from 15 participants, providing a substantial dataset for validating our methodology.
3) Evaluation of the impact of various time-frequency configurations, including continuous wavelet transform (CWT), pseudo Wigner-Ville distribution (PWVD), and short-time Fourier transform (STFT), on the performance of ViT architectures.
4) Fine-tuning and customization of the ViT model specifically for classifying in-ear PPG signals into distinct levels of mental stress.

The remainder of this article is organized as follows: Section II presents the methodology used in this study, and Section III presents the experimental results. In Section IV, we provide a comprehensive discussion of the proposed method. Finally, the article is concluded in Section V, which includes an outline of potential directions for future research.

II. Proposed Methodology

In this section, we present the proposed hardware architecture and the method used to classify mental stress levels based on in-ear PPG signals.

A. Proposed Hardware Architecture

This article introduces a hardware system specifically customized for stress monitoring. The core of this system is an in-ear PPG sensor paired with an accelerometer, designed to accurately measure and analyze physiological signals for stress detection. Central to the device is the MAX30102 PPG sensor by Maxim Integrated, a semiconductor company based in California, USA. This compact sensor (5.6 $\times$ 2.8 $\times$ 1.2 mm) features red (R) and infrared (IR) LEDs, operating at wavelengths of approximately $\lambda_R = 660$ nm and $\lambda_{IR} = 875$ nm, respectively. It is chosen for its small size, low power consumption (1.8 and 3.3 V for LEDs), programmable settings, and efficient ambient light rejection, making it ideal for wearable devices. The sensor, with a photodiode for light absorption measurement, captures data at a 16-bit resolution and a 100 Hz sampling rate.

Enhancing the system's capabilities, a six-axis MEMS inertial measurement unit (IMU) (MPU-6050) is integrated, combining a three-axis accelerometer and gyroscope. The MPU-6050's low power consumption, precision, and compact size (20 $\times$ 16 $\times$ 4 mm) make it highly suitable for wearable applications. Data acquisition is managed by an NRF5232 microcontroller from Nordic Semiconductor, featuring a 64-MHz Arm Cortex-M4 CPU, 512 KB flash, and 64 KB RAM. This microcontroller supports Bluetooth low energy for wireless data transmission and is powered by a 500 mAh, 3.7 V rechargeable lithium-ion polymer battery, selected for its capacity and compact dimensions. Additionally, the system includes power management circuitry, allowing for micro-USB charging.

To ensure efficient operation and safety, the hardware components, including the microcontroller, battery, and power management circuitry, are encased in a specially designed 3-D housing. The architecture and implementation of the proposed hardware system are illustrated in Fig. 1.

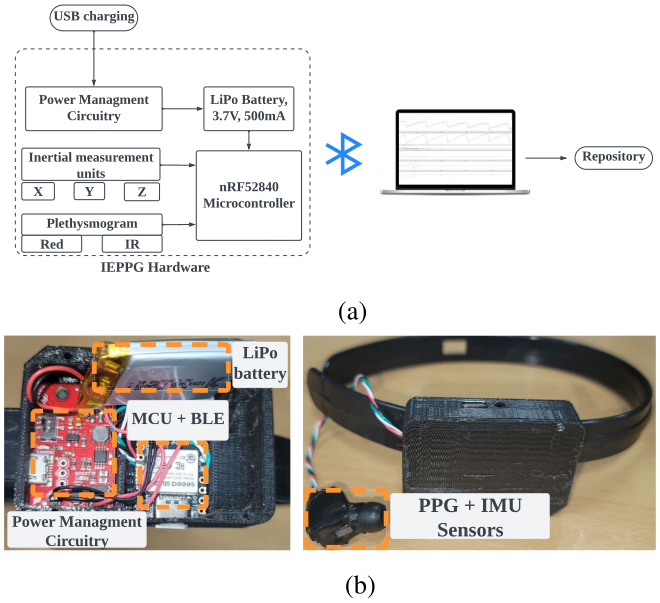


Fig. 1. Proposed hardware architecture and the implementation. (a) Proposed block diagram. (b) Implemented hardware.

TABLE I
Participant Characteristics

Study details	Study Participant Characteristics
Participants	15 (9 Males, 6 Females)
Age Range	23 to 37 years
BMI range	18.3 to 25.8 kg/m^2
Health status	Good
Exclusion criteria	Hypertension, cardiac arrhythmias, cognitive impairments, etc.

B. Proposed Mental Stress Level Classification System

This section will examine the proposed mental stress classification system designed to categorize stress into three levels. The comprehensive architecture of this method is depicted in Fig. 2.

1) Experimental Data Collection and Protocol: This stress exploration study involved 15 participants (nine males and six females). The relevant data were collected in a laboratory setting using an in-ear biosensor. The age of the participants ranged from 23 to 37 years, and they were in good health, maintaining a body mass index between 18.3 and 25.8 kg/m^2. Before experimenting, the participants were required to complete a detailed questionnaire about their cardiovascular and mental well-being, as well as disclose relevant medications that might influence the outcomes of the study. To ensure data integrity, individuals with hypertension, cardiac arrhythmias, or cognitive impairments were excluded. Please refer to Table I for a comprehensive summary of the study participants and their characteristics.

All participants performed a series of experiments, with each participant conducting three trials corresponding to different stress levels: low, medium, and high. These trials, lasting for three minutes each, were conducted within a controlled laboratory setting, adhering to the protocol delineated in Fig. 3. The study was structured to induce varying levels of

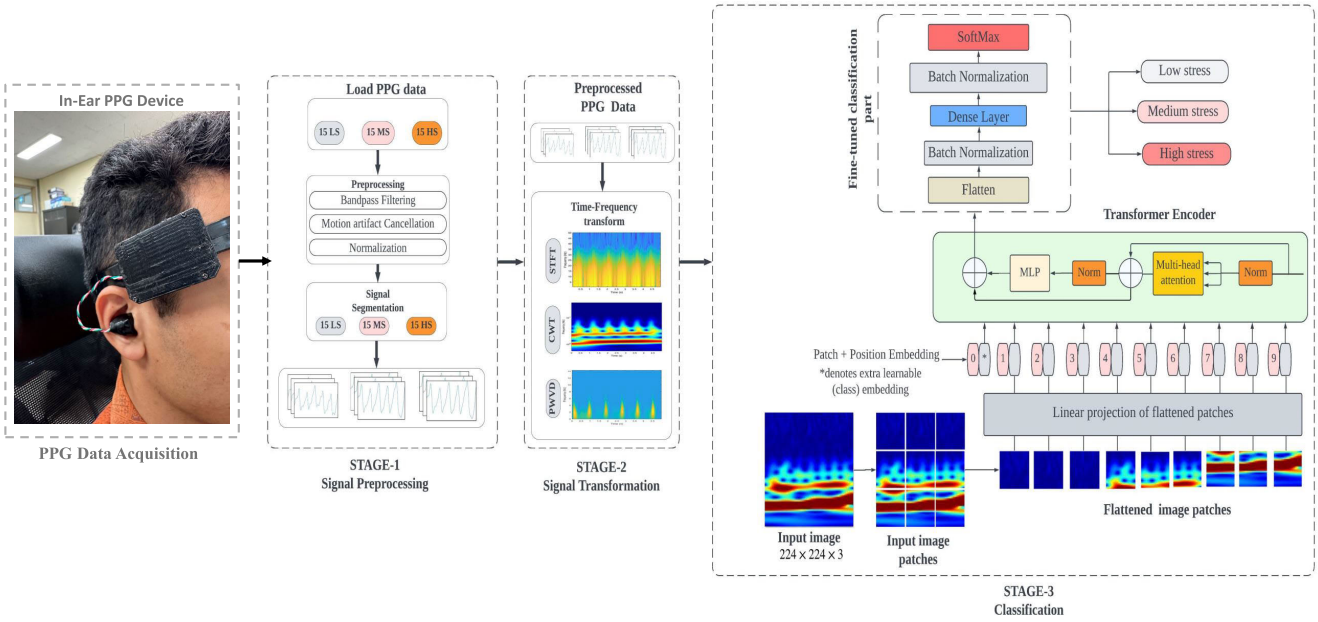


Fig. 2. Architecture of the proposed mental stress classification method.

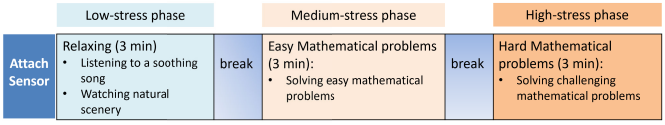


Fig. 3. Experimental protocol for stress-inducing activities involving mathematical problem solving and relaxation phases.

mental stress, particularly focusing on medium and high levels, through mental arithmetic tasks. For medium-stress induction, participants engaged in solving arithmetic involving two-digit numbers, incorporating operations such as addition, subtraction, and multiplication without calculators. For high stress, the complexity escalated to challenging problems, including three-digit number operations and multistep calculations, aiming to significantly increase cognitive load. Conversely, to induce low-stress conditions, participants were exposed to various relaxation techniques rather than cognitive tasks. These included listening to soothing music, watching serene natural scenery videos, or engaging in other calming activities designed to promote relaxation and a sense of peace. Each segment—whether aimed at stress induction or relaxation—lasted for three minutes, during which physiological data were continuously monitored at a sampling rate of 100 Hz, giving 18 000 data points per activity per participant. Fig. 4 visually depicts sample signals acquired under differing stress conditions, illustrating the physiological distinctions between low, medium, and high-stress states.

In addition to the behavioral induction of stress states, stress levels were assessed through self-reported questionnaires immediately following each task. These questionnaires were designed to measure the participants' subjective experiences and perceptions of stress, allowing for the classification into low, medium, or high-stress levels based on their responses. While the study did not incorporate biomarkers such as HRV or cortisol levels to gauge physiological stress responses, the use of self-reported measures provided valuable insights into the perceived stress levels of participants.

The study protocol underwent a thorough review by the institutional review board (IRB) and was approved under reference number 1041386-202305-HR-42-02. Furthermore, before starting the study, informed consent was obtained from all participants, and any required modifications to the original protocol were authorized by the IRB.

2) Noise Removal and Preprocessing: A fourth-order Chebyshev-II bandpass filter was used to clean up the raw PPG signals. It focused on frequencies between 0.5 and 10 Hz to improve signal quality by highlighting physiological frequencies and lowering noise. After that, we used an adaptive recursive least squares method on the raw PPG signals, using motion information from accelerometer data to eliminate motion artifacts [22]. Our analysis focused specifically on the axis of the accelerometer signal that displayed a strong correlation with the PPG signal, a relationship that the Pearson correlation coefficient quantified. Fig. 5 illustrates this process, showcasing both the raw and processed PPG signals postfiltering and motion artifact reduction, thereby providing insights into the efficacy of the applied preprocessing techniques.

3) Time-Frequency Analysis: In the proposed ViT model, the input data must be in the form of images. Thus, we transform the 1-D filtered PPG signals into 2-D image datasets using three time-frequency analysis techniques: CWT, PWVD, and STFT. The PPG signals, which are 3-min long for each subject, were initially segmented into 5-s nonoverlapping segments, resulting in 36 segments per subject and a total of 540 segments from 15 subjects. Despite the limited number of segments, the ViT model can effectively learn from this dataset due to its capacity to capture complex patterns from detailed

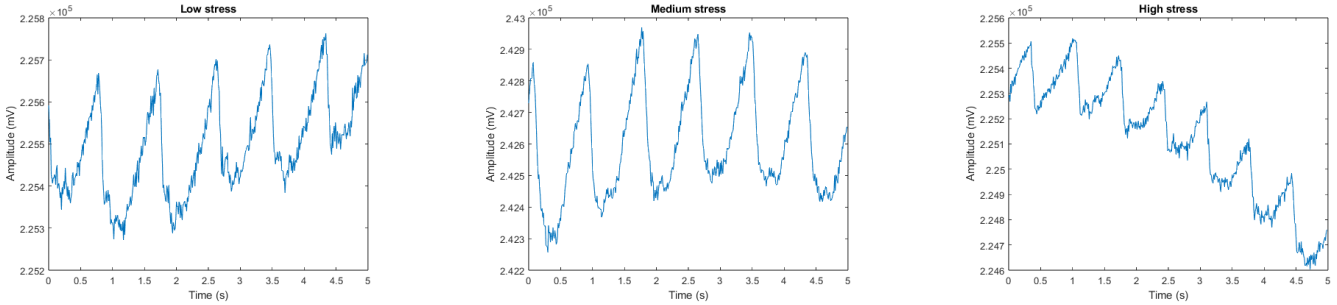


Fig. 4. Sample of raw PPG signals showing the mental stress status. (a) Low-stressed. (b) Medium-stressed. (c) High-stressed signals.

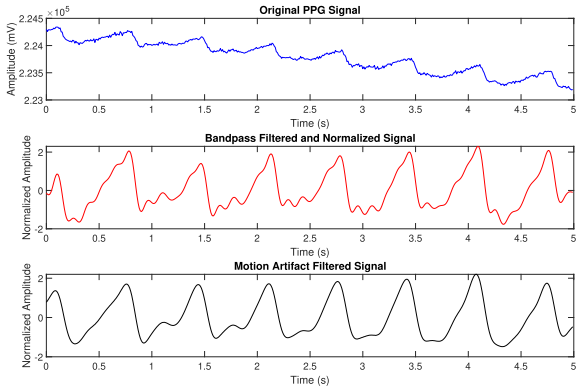


Fig. 5. Comparative analysis of PPG Signal processing techniques. The original PPG signal (top) is compared with the bandpass-filtered and normalized signal (middle) to enhance relevant frequencies. The bottom plot showcases the signal after motion artifact filtering, illustrating the effectiveness of the applied method in reducing noise and artifacts.

time-frequency representations. The resulting 224 × 224-pixel images provide sufficient information for training the model to identify mental stress patterns.

Additionally, we explored a sliding window approach with 50% overlap to generate overlapping segments. In this approach, consecutive 5-s segments share 2.5 s of data, resulting in 72 segments per subject and a total of 1080 segments across the dataset. Both nonoverlapping and 50% overlapping segmentations were applied to the PPG signals, and time-frequency transformations (CWT, PWVD, and STFT) were generated for each segment.

Sliding windows with overlap are generally considered to enhance classification performance by increasing the volume of data points and capturing subtle variations more effectively, particularly during transitions between mental states. However, our experiments indicate that nonoverlapping windows can achieve similar or slightly better classification accuracy while substantially reducing the computational effort and memory required for model training.

This exploration of both nonoverlapping and overlapping segmentation strategies allowed us to evaluate their potential for capturing stress-related patterns in the transformed time-frequency data. Both approaches provide structured input to the ViT model, ensuring compatibility with its architecture for classification.

a) Continuous wavelet transform: The CWT decomposes the PPG signal into a scalogram, aiding the ViT model in identifying mental stress. For a signal $x(t)$, the CWT is defined as

$$\mathrm{CWT}_x(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t)\omega\left(\frac{t-b}{a}\right) dt. \tag{1}$$

Here, a is the scale parameter, b the translation parameter, and $\omega(t)$ the wavelet function. The choice of the mother wavelet, such as the generalized Morse wavelet used in this study, significantly influences the quality of the scalogram images. The sample of the resulting PPG scalogram is shown in Fig. 6(b).

b) Pseudo Wigner-Ville distributions: The PWVD is used for a quantitative representation of signal energy in the time-frequency domain. The PWVD is defined as

$$\mathrm{PWVD}_x(t,\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} h(t)x\left(t+\frac{\tau}{2}\right) x^*\left(t-\frac{\tau}{2}\right) \times e^{-j\omega\tau} d\tau \tag{2}$$

where $h(t)$ is a sliding averaging window. The PWVD reduces the impact of cross-terms and provides time-frequency representations. The sample of the resulting PWVD spectrogram is shown in Fig. 6(c).

c) Short-time Fourier transform: The STFT analyzes both instantaneous frequency and amplitude variations within a signal [23]. For discrete digital signals, the STFT is given by

$$\mathrm{STFT}\{y[n]\} = Y(\alpha,\gamma) = \sum_{n=-\infty}^{\infty} y[n]s[n-\alpha]e^{-j\gamma n}. \tag{3}$$

The PPG signal, sampled at 100 Hz, was processed using a Hann window with a length of 128, and the resulting spectrograms were used for input to the ViT model. The sample of the resulting PPG spectrogram is shown in Fig. 6(d).

4) ViT Model: A transformer model is a type of neural network adept at understanding context and relationships in sequential data, such as the words within a sentence [24]. Employing attention mechanisms, particularly self-attention, transformers can discern complex interactions among distant elements in a sequence. This concept extends to image processing via ViT models, which interpret sequences of image patches, thereby broadening their applicability.

In the proposed methodology, the ViT model operates as follows.

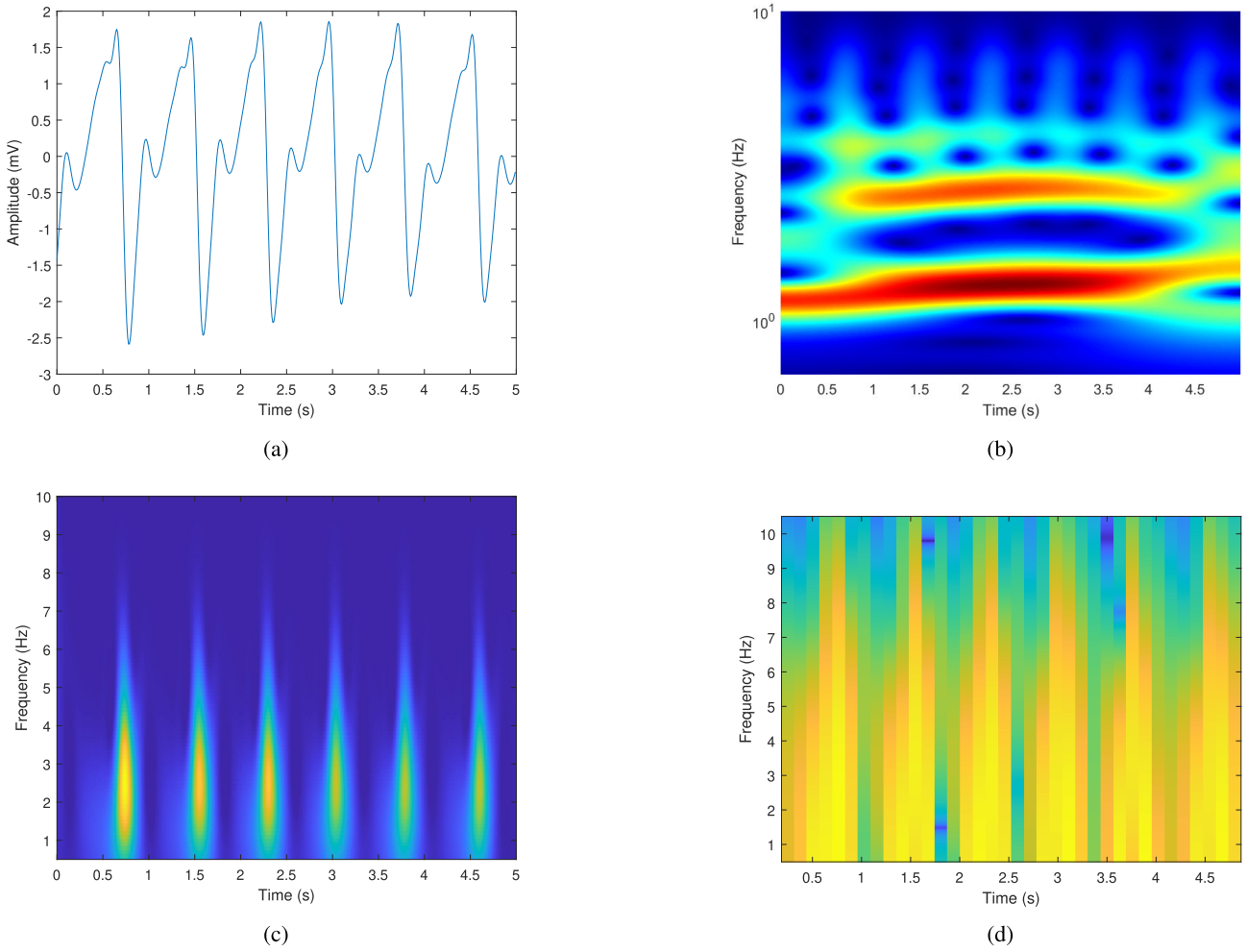


Fig. 6. Examples of (a) filtered signal, (b) CWT, (c) PWVD, and (d) STFT.

1) *Conversion of Images to Patches:* ViT models treat image patches akin to word tokens in NLP [24]. This approach, as suggested by Dosovitskiy et al. [25], converts an input image of size $H \times W$ into $N = (HW/P^2)$ patches, each of size $P \times P$.
2) *Flattening and Embedding of Patches:* Each patch is flattened into a vector X_{np} of length $P^2 \times C$ for $n = 1, \ldots, N$. These vectors are then mapped to D dimensions through a linear projection involving an embedding matrix E, initially randomized [24].
3) *Learnable Embedding and Positional Embedding:* The flattened patches, represented as embedded vectors, are given by the equation

$$z_0 = \left[X_{\text{class}}; X_{1p}E; \ldots; X_{Np}E\right] + E_{\text{pos}}. \quad (4)$$

Here, X_{class} incorporates both trainable embeddings for classification and positional embeddings to maintain the sequential order of patches.
4) *MLP Head:* The transformer encoder outputs are directed to an multilayer perceptron (MLP) head for classification. The MLP primarily focuses on the output associated with the class embedding, X_{class}, producing a probability distribution for image labels.

For this study, we employed three distinct ViT models: a base model with 16 × 16 input patch size (ViT-b16), a base model with 32 × 32 input patch size (ViT-b32), and a large model with 32 × 32 input patch size (ViT-l32). While ViT-b16 and ViT-b32 are considered "base" models with fewer layers and parameters, designed for general purposes, the ViT-l32 is classified as a "large" model due to its increased depth and width, allowing for more complex feature representation. These models were used to identify cases of low, mild, and high stress, aiming to evaluate their effectiveness in classifying mental stress compared to other scenarios.

In the pretraining phase, the MLP serves as the classification head, which is substituted with a dedicated classification component during the fine-tuning phase. As shown in Fig. 7, we implement a series of layers to each ViT model, including a flatten layer, a batch normalization layer, and a dense layer with 32 units, followed by an additional batch normalization layer. Moreover, the SoftMax function is employed to derive the classification probabilities for different mental stress levels (low, medium, or high).

5) K-Fold Cross-Validation: To mitigate the risk of overfitting and to ensure a robust evaluation, we employed a between-subjects fivefold cross-validation approach in the assessment of the proposed method. Specifically, the dataset was partitioned into five equally sized subsets, with care taken to ensure that data from any single subject were exclusively allocated to one subset only. Consequently, each fold was

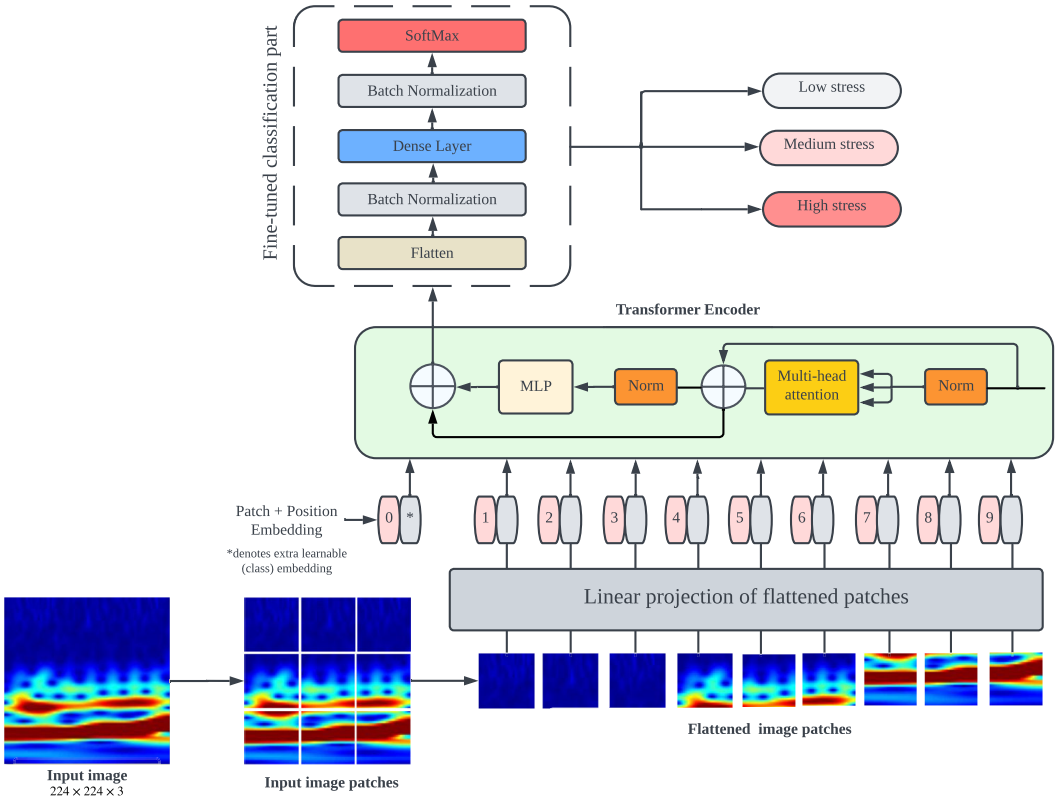


Fig. 7. Block diagram of the proposed ViT-based mental stress classification method.

utilized once as the testing set, while the remaining folds formed the training set, ensuring that subjects' data did not overlap between these sets. This approach helped in preventing potential information leakage and maintained the integrity of the evaluation process by guaranteeing that the performance metrics were based on the model's ability to generalize across unseen subjects. This rigorous validation framework is aimed at providing dependable and representative performance metrics, reflecting the true efficacy of the proposed method in diverse scenarios.

6) Hyperparameter Configuration: We implemented specific hyperparameter configurations to effectively train the ViT models for the mental stress classification task. In this process, the models were trained for 30 epochs using the Adam optimizer, with a learning rate set at 1×10^{-4}. We adopted a batch size of 64 to enhance the training procedure and introduced exponential decay. For data partitioning, we followed a ratio of 7:2:1 for the distribution of training, validation, and testing data, respectively. This allocation ensured an ample amount of data for training while facilitating robust evaluation. To boost performance and generalization, the ViT models employed the GELU activation function, recognized for its effectiveness in neural networks, as well as an $L2$ regularizer to mitigate overfitting. These meticulously chosen hyperparameters (see Table II) played a pivotal role in training ViT models and achieving optimal results in mental stress classification.

7) Performance Evaluation: To assess the performance of the proposed mental stress classification method, we conducted a rigorous evaluation using well-established metrics, i.e., accuracy, precision, recall, $F1$-score, and area under the receiver operating curve (AUC) to provide a comprehensive understanding of the effectiveness of the proposed model. The AUC, which is calculated through the receiver operating characteristic (ROC) curve, is an important performance metric that is frequently used in medical classification challenges. The AUC metric emphasizes the balance between the correct and incorrect classifications obtained by the model.

TABLE II
HYPERPARAMETER CONFIGURATION FOR VIT MODEL

Hyperparameter	Value
Number of Epochs	30
Optimizer	Adam
Learning Rate	1e-4
Batch Size	64
Learning Rate Decay	Exponential Decay
Data Split (Train:Validation:Test)	7:2:1
Activation Function	GELU
Regularization	L2 Regularizer

III. EXPERIMENTAL RESULTS AND ANALYSIS

A key aspect of this study is the innovative transformation of in-ear PPG signals into time-frequency images (CWT, PWVD, and STFT) and the subsequent use of ViT models for classification. This unique approach harnesses the image-processing capabilities of ViT, setting it apart from traditional machine-learning methods that typically analyze raw or minimally processed physiological signals. This foundation allows for a more detailed analysis of stress levels

TABLE III
CLASSIFICATION RESULTS OF VIT MODELS WITH DIFFERENT TIME-FREQUENCY METHODS (USING NONOVERLAPPING SLIDING WINDOW). (a) CWT METHOD. (b) PWVD METHOD. (c) STFT METHOD

(a)

Metrics (%)	ViT-b16	ViT-b32	ViT-l32
Accuracy	95.86 ± 1.87	96.05 ± 2.20	96.30 ± 3.30
Precision	95.94 ± 1.83	96.21 ± 2.11	96.30 ± 3.34
Recall	95.86 ± 1.87	96.05 ± 2.20	96.30 ± 3.30
F1-score	95.85 ± 1.89	96.07 ± 2.19	96.27 ± 3.34
AUC	0.9868 ± 0.0022	0.9849 ± 0.0043	0.9775 ± 0.0084

(b)

Metrics (%)	ViT-b16	ViT-b32	ViT-l32
Accuracy	91.73 ± 3.28	79.69 ± 6.32	85.93 ± 4.10
Precision	92.36 ± 2.67	81.12 ± 4.61	86.59 ± 4.07
Recall	91.73 ± 3.28	79.69 ± 6.32	85.93 ± 4.10
F1-score	91.68 ± 3.36	79.57 ± 6.45	85.89 ± 4.09
AUC	0.9694 ± 0.0010	0.9197 ± 0.0013	0.9660 ± 0.0015

(c)

Metrics (%)	ViT-b16	ViT-b32	ViT-l32
Accuracy	97.78 ± 2.10	97.16 ± 1.91	97.47 ± 1.49
Precision	97.82 ± 2.02	97.26 ± 1.78	97.51 ± 1.45
Recall	97.78 ± 2.10	97.16 ± 1.91	97.47 ± 1.49
F1-score	97.79 ± 2.08	97.16 ± 1.91	97.46 ± 1.50
AUC	0.9886 ± 0.0059	0.9901 ± 0.0051	0.9867 ± 0.0042

TABLE IV
CLASSIFICATION RESULTS OF VIT MODELS WITH DIFFERENT TIME-FREQUENCY METHODS (USING OVERLAPPING SLIDING WINDOW). (a) CWT METHOD. (b) PWVD METHOD. (c) STFT METHOD

(a)

Metrics (%)	ViT-b16	ViT-b32	ViT-l32
Accuracy	95.12 ± 4.44	96.17 ± 3.22	96.03 ± 2.42
Precision	95.19 ± 4.35	96.50 ± 3.95	96.04 ± 2.62
Recall	95.12 ± 4.44	96.17 ± 3.22	96.03 ± 2.42
F1-score	95.11 ± 4.44	96.10 ± 3.30	96.02 ± 2.79
AUC	0.9875 ± 0.0121	0.9877 ± 0.0113	0.9863 ± 0.0112

(b)

Metrics (%)	ViT-b16	ViT-b32	ViT-l32
Accuracy	92.58 ± 3.50	89.30 ± 2.48	92.11 ± 3.07
Precision	92.66 ± 3.42	89.64 ± 2.41	92.37 ± 2.93
Recall	92.58 ± 3.50	89.30 ± 2.48	92.11 ± 3.07
F1-score	92.60 ± 3.48	89.37 ± 2.43	92.16 ± 3.01
AUC	0.9883 ± 0.0037	0.9747 ± 0.0073	0.9893 ± 0.0024

(c)

Metrics (%)	ViT-b16	ViT-b32	ViT-l32
Accuracy	95.07 ± 1.89	96.14 ± 1.48	96.96 ± 2.16
Precision	95.46 ± 1.74	96.30 ± 1.34	97.12 ± 2.02
Recall	95.07 ± 1.89	96.14 ± 1.48	96.96 ± 2.16
F1-score	95.13 ± 1.84	96.16 ± 1.45	96.97 ± 2.03
AUC	0.9801 ± 0.0193	0.9879 ± 0.0075	0.9873 ± 0.0084

by leveraging ViT's advanced feature extraction capabilities, contributing to enhanced classification accuracy.

The system modules employed in this study were executed on a workstation running the 64-bit Windows 10 Pro operating system. This workstation was powered by an AMD Ryzen 5 5600G processor operating at a clock speed of 3.90 GHz and equipped with Radeon Graphics, accompanied by 16 GB of RAM. For training the ViT model and performing the fivefold cross-validation, an NVIDIA RTX 3090 GPU was used.

A. Performance Comparison of Various Time-Frequency Methods on ViT Models

In this section, we compare the classification performance of three fine-tuned ViT models (ViT-b16, ViT-b32, and ViT-l32) across different time-frequency methods: CWT, PWVD, and STFT, using both nonoverlapping and overlapping sliding windows. The results are summarized in Table III (nonoverlapping) and Table IV (overlapping), while the corresponding ROC curves and AUC values are shown in Figs. 8 and 9.

The CWT method revealed that the ViT-l32 model demonstrated superior performance across both configurations. In the nonoverlapping setup [Table III(a)], the ViT-l32 model achieved an accuracy of 96.30%, with corresponding precision, recall, and $F1$-score values of 96.30%, 96.30%, and 96.27%, respectively. However, in the overlapping window configuration [Table IV(a)], the accuracy slightly decreased to 96.03%, while the AUC improved from 0.9775 to 0.9983 [Fig. 9(a)]. This suggests that the overlapping configuration enhances the model's ability to discriminate between stress levels, even with a minor reduction in accuracy. The ViT-b16 model performed consistently well, achieving 95.12% accuracy and an AUC of 0.9975 with overlapping windows. Meanwhile, the ViT-b32 model exhibited a drop in accuracy to 89.17% with overlapping windows, though its precision remained stable at 89.50%. These results suggest that overlapping windows provide finer temporal resolution, enhancing performance for models like ViT-l32, though they may introduce accuracy trade-offs for others like ViT-b32.

The PWVD method showed that overlapping sliding windows significantly boosted the performance of the ViT models. The ViT-b16 model, which achieved 91.73% accuracy in the nonoverlapping setup [Table III(b)], improved to 92.58% with overlapping windows [Table IV(b)], along with a high AUC of 0.9883 [Fig. 9(b)]. Similarly, the ViT-l32 model's accuracy increased from 85.93% to 92.11%, demonstrating that PWVD benefits greatly from overlapping windows. Notably, the ViT-b32 model experienced the most substantial improvement, with its accuracy increasing from 79.69% to 89.30%. These results highlight the effectiveness of overlapping windows in extracting subtle time-frequency patterns, particularly for more complex models like ViT-l32 and ViT-b32.

The STFT method showed strong overall performance across configurations. In the nonoverlapping setup [Table III(c)], all models performed well, with the ViT-b32 model achieving the highest balance of metrics, making it the best-performing model in this study. Specifically, the ViT-b32 model attained an accuracy of 97.16%, with high precision, recall, and $F1$-score values, underscoring its robustness for stress level classification. In comparison, the ViT-b16 and ViT-l32 models achieved accuracies of 97.78% and 97.47%, respectively, also performing effectively but not surpassing the consistency observed with ViT-b32. The ROC curve for this configuration in Fig. 8(c) illustrates the high AUC values, with ViT-b32 reaching an AUC of 0.9901.

In the overlapping configuration [Table IV(c)], the accuracy of the ViT-l32 model dropped slightly to 96.96%, though its AUC improved from 0.9867 to 0.9873 [Fig. 9(c)]. This suggests that overlapping windows allowed the model to better

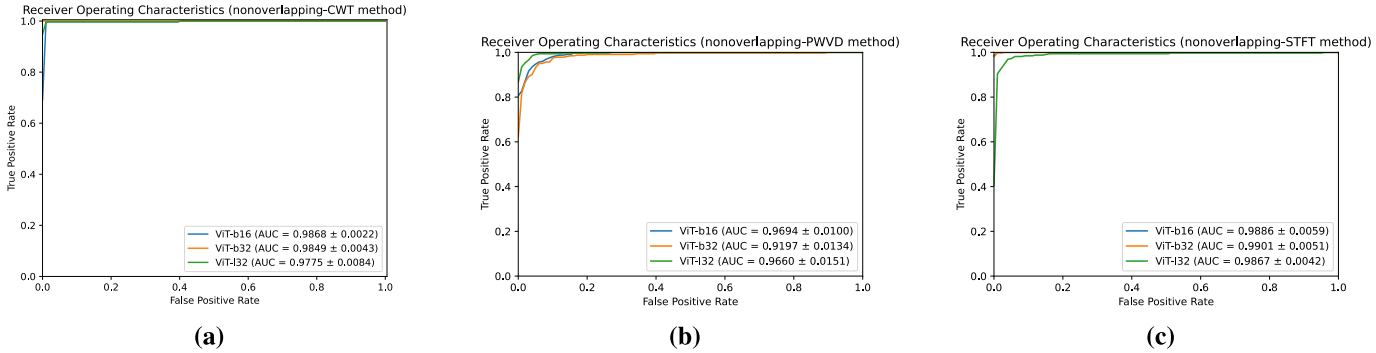


Fig. 8. Average ROC curves of fivefold cross-validation for different signal processing methods (nonoverlapping sliding windows). (a) CWT. (b) PWVD. (c) STFT.

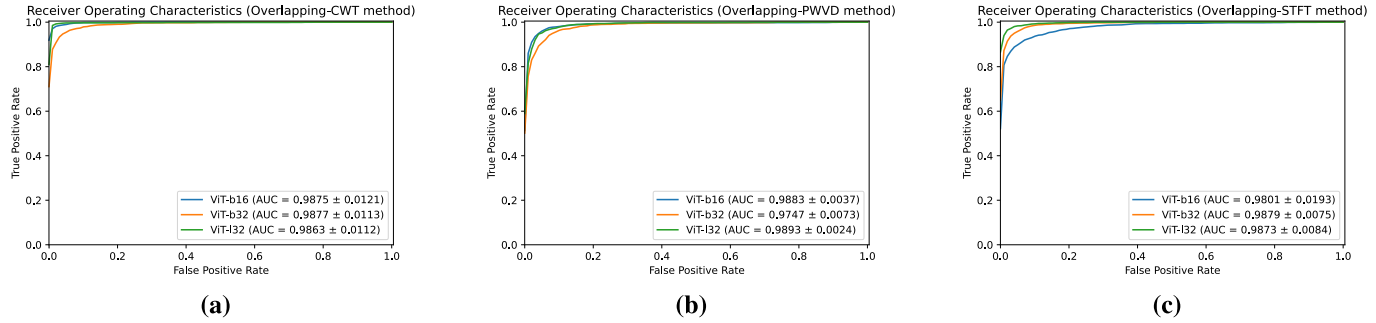


Fig. 9. Average ROC curves of fivefold cross-validation for different signal processing methods (overlapping sliding windows). (a) CWT. (b) PWVD. (c) STFT.

distinguish between stress and nonstress levels, even with a small reduction in accuracy. The ViT-b32 model showed a modest change in accuracy with overlapping windows, from 97.16% to 96.14%, while achieving the highest AUC in this setup, with a value of 0.9879 [Fig. 9(c)]. The ViT-b16 model, with an AUC of 0.9801, showed stable performance in terms of accuracy, achieving 95.07%.

The results demonstrate that the nonoverlapping sliding window configuration with the ViT-b32 model and STFT method provides an optimal balance of high accuracy and stability, making it the preferred choice for stress classification. Although overlapping sliding windows generally enhance AUC—a key metric in stress detection for capturing subtle variations in time-frequency patterns—the nonoverlapping setup with ViT-b32 still achieves high AUC and accuracy, especially with the STFT method. Notably, ViT-l32 also performed well with CWT and STFT methods, effectively capturing complex patterns; however, the ViT-b32 model proved particularly resilient, showing robust classification performance across different time-frequency methods, especially with STFT in both overlapping and nonoverlapping configurations. This combination of the ViT-b32 model with the STFT method in a nonoverlapping setup underscores the importance of carefully selecting model architecture, time-frequency method, and window configuration to maximize the accuracy and reliability of mental stress detection systems.

B. Time Complexity Analysis and Model Efficiency

We analyzed the training time for ViT-b16, ViT-b32, and ViT-l32 models using three signal processing methods (CWT, PWVD, and STFT) with both nonoverlapping and overlapping sliding windows (Figs. 10 and 11). The results show significant variations in time complexity based on both the model and the signal processing method used.

In the nonoverlapping sliding window configuration (Fig. 10), the ViT-b16 model consistently exhibited the highest time complexity across all methods. In contrast, the ViT-b32 model demonstrated the lowest time complexity, showing considerable efficiency, particularly with CWT and STFT. The ViT-l32 model's time complexity was moderate compared to the other two models, indicating a balance between computational cost and efficiency across methods. These results indicate that, for nonoverlapping windows, the ViT-b32 model is the most efficient across all signal processing methods, while the ViT-b16 model incurs the highest computational cost.

In the overlapping sliding window configuration (Fig. 11), the training time for all models increased across methods due to the additional computational load introduced by finer temporal segmentation. The ViT-b16 model remained the most time-consuming, especially with STFT. While still efficient, the ViT-b32 model showed an increase in time complexity but maintained its lead as the most efficient among the three. The ViT-l32 model demonstrated intermediate time complexity, reflecting the additional burden of overlapping windows while balancing efficiency and computational demands.

These results underscore the importance of considering both model architecture and signal processing method when optimizing for time complexity. Across both configurations, the ViT-b32 model consistently emerged as the most efficient, while the ViT-b16 model exhibited the highest time

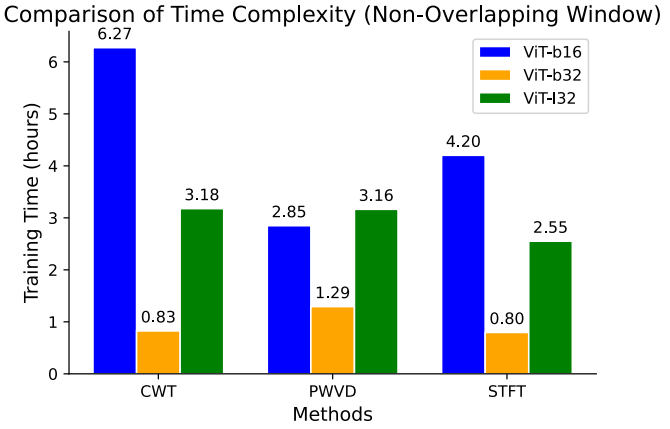


Fig. 10. Time complexity of CWT, PWVD, and STFT methods for each ViT model (nonoverlapping sliding window).

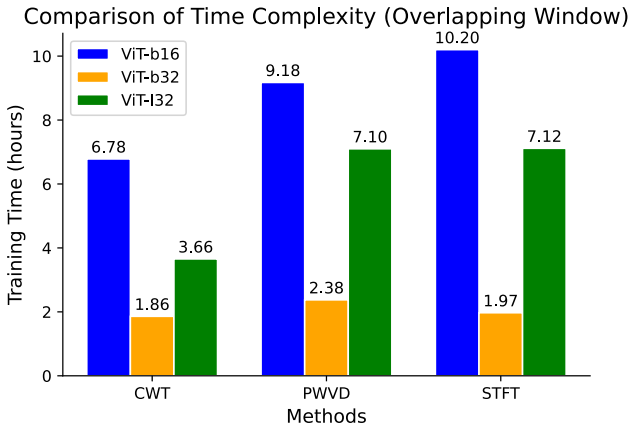


Fig. 11. Time complexity of CWT, PWVD, and STFT methods for each ViT model (overlapping sliding window).

complexity. This analysis illustrates how overlapping windows, while beneficial for capturing more detailed temporal patterns, also lead to a significant increase in training time.

IV. Discussion

This study leveraged in-ear PPG signals to explore the effectiveness of various ViT models for detecting and classifying mental stress at multiple levels. By integrating these unique in-ear PPG signals with advanced deep learning techniques, we gained meaningful insights into the nuanced dynamics of stress markers. An extensive analysis was conducted on different ViT models using three time-frequency methods—CWT, PWVD, and STFT—each tested with both nonoverlapping and overlapping sliding window configurations.

Applying the CWT method to in-ear PPG signals, the ViT-l32 model exhibited strong performance in both configurations, achieving high accuracy and AUC values, especially in the overlapping setup [Table IV(a), Fig. 9(a)]. However, the ViT-b32 model demonstrated a favorable balance between accuracy and efficiency, particularly in the nonoverlapping configuration, with significantly reduced training time (Fig. 10). These results suggest that the ViT-b32 model paired with CWT is ideal for applications prioritizing efficiency, while the ViT-l32 model is better suited for tasks that require high accuracy and fine-grained classification.

The STFT method demonstrated robust overall performance. In the nonoverlapping configuration [Table III(c)], the ViT-b32 model achieved a high accuracy of 97.16% with a strong AUC of 0.9901 [Fig. 8(c)], making it the most balanced performer in this setup. STFT's capacity to provide stable time and frequency resolution allowed it to capture temporal dynamics and frequency characteristics of stress markers effectively. Although the ViT-b16 model showed comparable accuracy, the ViT-b32 model's performance makes it a promising choice for applications requiring both accuracy and efficiency in a nonoverlapping setup. In the overlapping configuration [Table IV(c)], the ViT-b32 model maintained strong performance with an AUC of 0.9879 [Fig. 9(c)], indicating that overlapping windows enhance STFT's ability to capture subtle temporal patterns, albeit with an increase in computational load.

With the PWVD method, the use of overlapping windows consistently enhanced model performance, particularly for the ViT-b32 model, which experienced a substantial accuracy increase from the nonoverlapping setup [Tables III(b) and IV(b)]. The ViT-l32 model also benefited from overlapping windows, underscoring the importance of finer temporal segmentation when analyzing nonstationary signals like PPG. Despite these enhancements, the ViT-b16 model retained its advantage in AUC [Fig. 9(b)], suggesting its suitability for applications where discriminative capacity is a priority.

These findings reveal the nuanced performance variations among ViT models in stress detection tasks, influenced by the choice of signal processing method and the specific properties of in-ear PPG signals. The ViT-b32 model stands out for its efficiency, particularly when combined with CWT and STFT in a nonoverlapping configuration, making it a robust candidate for real-time applications. Conversely, the ViT-l32 model, with its high accuracy and AUC in the overlapping configuration, is well-suited for scenarios where detailed and precise classification is paramount. This study underscores the critical role of model selection, time-frequency methods, and window configuration in optimizing mental stress detection systems.

A. Comparative Analysis of Fine-Tuned ViT-b32 and Other Deep Learning Models for Stress Level Classification

Comprehensive comparative analyses were conducted between the proposed ViT-b32 model and commonly used deep learning models, including VGG19, DenseNet169, DenseNet201, InceptionResNetV2, InceptionV3, and YOLOv8 (You Only Look Once), to evaluate their effectiveness in mental stress level classification. As presented in Table V and Fig. 12, the proposed ViT-b32 model achieved superior performance compared to the pretrained CNN models, with an accuracy of 97.78%, precision of 97.82%, recall of 97.78%, and $F1$-score of 97.79%. These metrics reflect the robustness of the ViT-b32 model in accurately identifying stress levels, underscoring its potential for applications in stress management and mental health monitoring. The transformation of PPG signals into time-frequency images and

TABLE V
COMPARISON OF PROPOSED VIT MODEL WITH OTHER STATE-OF-THE-ART DEEP LEARNING MODELS

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC
VGG16	86.67 ± 15.57	85.38 ± 19.00	86.67 ± 15.57	84.92 ± 19.10	0.9795 ± 0.00553
DenseNet169	96.23 ± 2.82	96.29 ± 2.78	96.23 ± 2.82	96.24 ± 2.81	0.9834 ± 0.0031
DenseNet201	96.60 ± 2.72	96.69 ± 2.59	96.60 ± 2.72	96.62 ± 2.69	0.9868 ± 0.0023
InceptionResNetV2	92.10 ± 6.81	92.99 ± 5.35	92.10 ± 6.81	92.13 ± 6.72	0.9770 ± 0.0055
InceptionNetV3	95.19 ± 4.14	95.20 ± 4.17	95.19 ± 4.14	95.17 ± 4.16	0.9713 ± 0.0039
YOLOv8n-cls	96.73 ± 1.18	96.76 ± 1.16	96.73 ± 1.18	96.75 ± 1.21	0.9861 ± 0.0067
ViT-b32 (STFT, NO)	**97.78 ± 2.10**	**97.82 ± 2.02**	**97.78 ± 2.10**	**97.79 ± 2.08**	**0.9901 ± 0.0051**

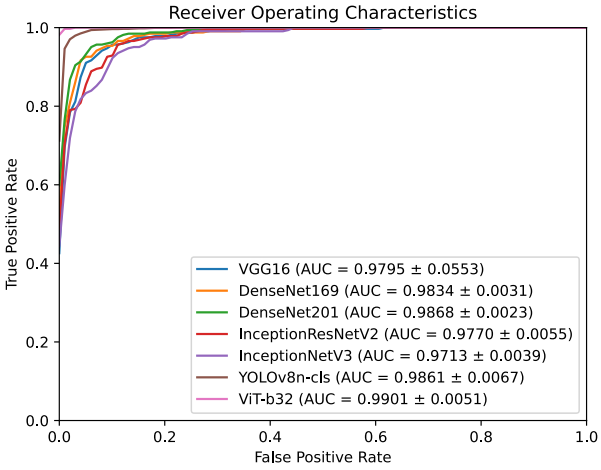


Fig. 12. Comparison of average ROC curves of the proposed ViT model with other state-of-the-art deep learning models.

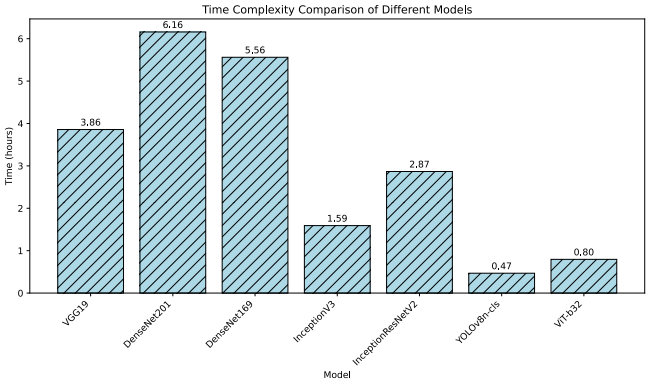


Fig. 13. Time complexity comparison of the VGG19, DenseNet201, DenseNet169, InceptionV3, InceptionResNetV2, YOLOv8, and proposed ViT-b32 methods.

the application of ViT for classification offers a significant improvement over previous machine-learning approaches. By leveraging the advanced feature extraction capabilities of ViT, this method enhances classification accuracy, demonstrating the benefit of using image-based deep learning models for stress detection. The results demonstrate the advantage of leveraging innovative architectures like ViT-b32 for enhanced performance in stress classification tasks, especially in scenarios where high accuracy and reliability are critical.

In terms of computational efficiency, Fig. 13 shows a comparison of training times across different models. The YOLOv8 model demonstrated the lowest time complexity at 0.47 h, making it the most time-efficient choice for resource-constrained environments. The ViT-b32 model also proved to be efficient, with a time complexity of 0.80 h, which is still significantly lower than many CNN counterparts, such as DenseNet201 with the highest computational cost of 6.16 h. Although YOLOv8 is more time-efficient, the ViT-b32 model offers a superior balance between accuracy and efficiency, outperforming YOLOv8 in accuracy metrics, which are essential for stress detection tasks that demand high precision and reliability.

These results provide valuable insights for researchers and practitioners when choosing models for specific applications. The high performance of the ViT-b32 model and the efficiency of YOLOv8 highlight the trade-offs between accuracy and time complexity. While ViT-b32 is suitable for real-time stress detection tasks where accuracy is paramount, YOLOv8 may be an ideal choice in scenarios where computational resources are highly limited. Future research could explore optimization strategies tailored to different models, potentially leading to even more refined choices for deep learning applications in stress management.

B. Comparison With State-of-the-Art Stress Classification Methods

In evaluating our approach against contemporary mental stress classification techniques, our method, utilizing the ViT-b32 model with PPG data, has demonstrated significant advancement in classification accuracy (see Table VI). When employing a fivefold cross-validation (fivefold CV) approach, our technique achieved an impressive 97.78% accuracy across three classes of mental stress. This performance not only attests to the efficacy of our approach but also highlights the potential of the ViT model in complex stress classification tasks.

To ensure a fair and rigorous comparison with existing studies, we extended our validation approach to include leave-one-subject-out cross-validation (LOSO CV), a method widely regarded for its stringent testing conditions in personalized health monitoring. Under the LOSO CV protocol, our method maintained a robust accuracy of 96.99%, demonstrating its effectiveness and generalizability across different individuals—an essential factor in real-world applications.

This LOSO CV result is particularly noteworthy when compared to other prominent methods that have employed similar validation strategies. For instance, Zubair and Yoon [15] achieved a 94.33% accuracy using an SVM classifier with

TABLE VI
PERFORMANCE COMPARISON OF THE PROPOSED METHOD WITH THE STATE-OF-THE-ART METHODS

Authors	Classifiers	Validation protocol	Results (Accuracy)
Zubair *et al.* [15]	SVM	LOSO CV	94.33%
Heo *et al.* [16]	Ensemble classifier	LOSO CV	96.50%
Zangróniz *et al.* [19]	Discriminant tree-based model	10-fold stratified CV	82.35%
Kalra *et al.* [18]	DNN	10-fold CV	91%
Barki *et al.* [20]	CNN	-	96.02%
Our method (LOSO CV)	ViT-b32	LOSO CV	**96.99%**
Our method (5-fold CV)	ViT-b32	5-fold CV	**97.78%**

LOSO CV. Similarly, Heo et al. [16] utilized an Ensemble classifier under the same LOSO CV framework, reaching a 96.50% accuracy. These comparisons underscore the superior performance of our method within the stringent testing framework provided by LOSO CV.

Moreover, when compared to other methods utilizing different cross-validation strategies, such as the 82.35% accuracy attained by Mukherjee et al. [19] using a tenfold stratified CV with a Discriminant tree-based model, or the 91% accuracy reported by Kalra and Sharma [18] under a tenfold CV with a DNN approach, our method continues to demonstrate its leading edge. Additionally, Barki and Chung [20], who reported a 96.02% accuracy using a CNN model, highlights the competitive performance of our approach even without specifying their validation protocol.

Overall, the exceptional performance of our ViT-b32 model, verified through both fivefold CV and LOSO CV protocols, confirms its high accuracy and effective handling of diverse mental stress categories. These findings strengthen the case for adopting ViT models in the biometric analysis of stress, setting a new benchmark in the field and providing a strong foundation for future research and applications in stress classification.

Generally, we introduced an innovative approach by integrating ViT models with in-ear PPG for enhanced mental stress detection. This combination has shown improved accuracy in stress classification, marking a significant advancement in the field. However, our research is not without limitations. The primary constraint is the limited sample size, which might affect the generalizability of our findings. Furthermore, the study's focus on a controlled environment may limit the applicability of the results to more dynamic or physically active real-world scenarios. Despite these limitations, our study offers substantial contributions to mental health monitoring. The potential of this method for real-time stress monitoring in various settings, coupled with its implications for mental health interventions, highlights its significance. Future work will aim to address these limitations, expanding the scope and applicability of our findings in broader contexts.

V. CONCLUSION

In this study, we conducted a comprehensive investigation into the use of ViT models with a novel in-ear PPG device for detecting and classifying mental stress across multiple levels, utilizing various time-frequency transform methods. The results demonstrated that the proposed ViT-b32 model, especially when paired with the STFT method in a nonoverlapping configuration, represents a significant advancement in mental stress classification. This combination achieved outstanding accuracy, precision, recall, $F1$-score, and inference time, consistently outperforming conventional models with an average accuracy of 97.78%. The unique in-ear PPG device, integrated with the ViT model architecture, enabled the capture of fine-grained physiological signals and allowed the model to effectively leverage these distinct features for accurate stress classification.

Additionally, our analysis highlighted that while the ViT-b32 model provides an optimal balance between accuracy and efficiency, other models like YOLOv8 offer remarkable time efficiency, presenting alternative solutions for applications with limited computational resources. This nuanced performance comparison underscores the importance of selecting models based on specific application requirements, whether prioritizing accuracy, efficiency, or both.

For future work, several directions are proposed to enhance the practicality and usability of this system. A primary focus will be on miniaturizing the in-ear PPG device and integrating it into compact, wearable form factors such as earphones or earbuds. This approach would allow for discreet, continuous, and real-time stress monitoring in daily life without requiring additional bulky equipment. Additionally, efforts will focus on employing a lightweight model that is efficient enough for on-board processing within these wearable devices. This will involve using model compression techniques, such as pruning and quantization, to reduce computational load and memory requirements, enabling the model to function effectively on low-power, embedded hardware.

REFERENCES

[1] R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters, "Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication," *Arch. Gen. Psychiatry*, vol. 62, no. 6, p. 617, Jun. 2005, doi: 10.1001/archpsyc.62.6.617.

[2] (2023). *Stress*. Accessed: Aug. 1, 2023. [Online]. Available: https://www.apa.org/topics/stress

[3] E. J. Vella, "Psychosocial factors in coronary heart disease," in *The Wiley Encyclopedia of Health Psychology*. Hoboken, NJ, USA: Wiley, Sep. 2020, pp. 529–535, doi: 10.1002/9781119057840.CH103.

[4] S. C. Segerstrom and G. E. Miller, "Psychological stress and the human immune system: A meta-analytic study of 30 years of inquiry," *Psychol. Bull.*, vol. 130, no. 4, pp. 601–630, Jul. 2004, doi: 10.1037/0033-2909.130.4.601.

[5] S. Cohen, J. D. Denise, and G. E. Miller, "Psychological stress and disease," *J. Amer. Med. Assoc.*, vol. 298, no. 14, pp. 1685–1687, Oct. 2007, doi: 10.1001/jama.298.14.1685.

[6] T. A. Revenson, K. Kayser, and G. Bodenmann, Eds., *Couples Coping With Stress: Emerging Perspectives on Dyadic Coping*. Washington, DC, USA: American Psychological Association, 2005, doi: 10.1037/11031-000.

[7] P. E. Greenberg et al., "The economic burden of adults with major depressive disorder in the United States (2010 and 2018)," *PharmacoEconomics*, vol. 39, no. 6, pp. 653–665, Jun. 2021, doi: 10.1007/S40273-021-01019-4.
[8] X. Dai and Y. Ding, "Mental health monitoring based on multiperception intelligent wearable devices," *Contrast Media Mol. Imag.*, vol. 2021, pp. 1–7, Nov. 2021, doi: 10.1155/2021/8307576.
[9] A. Goyal, S. Singh, D. Vir, and D. Pershad, "Automation of stress recognition using subjective or objective measures," *Psychol. Stud.*, vol. 61, no. 4, pp. 348–364, Dec. 2016, doi: 10.1007/S12646-016-0379-1.
[10] U. M. Nater and N. Rohleder, "Salivary alpha-amylase as a non-invasive biomarker for the sympathetic nervous system: Current state of research," *Psychoneuroendocrinology*, vol. 34, no. 4, pp. 486–496, May 2009, doi: 10.1016/j.psyneuen.2009.01.014.
[11] S. S. Dickerson and M. E. Kemeny, "Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research," *Psychol. Bull.*, vol. 130, no. 3, pp. 355–391, 2004, doi: 10.1037/0033-2909.130.3.355.
[12] P. H. Charlton, P. Celka, B. Farukh, P. Chowienczyk, and J. Alastruey, "Assessing mental stress from the photoplethysmogram: A numerical study," *Physiol. Meas.*, vol. 39, no. 5, May 2018, Art. no. aabe6a, doi: 10.1088/1361-6579/aabe6a.
[13] P. Celka, P. H. Charlton, B. Farukh, P. Chowienczyk, and J. Alastruey, "Influence of mental stress on the pulse wave features of photoplethysmograms," *Healthcare Technol. Lett.*, vol. 7, no. 1, pp. 7–12, Feb. 2020, doi: 10.1049/htl.2019.0001.
[14] Z.-H. Wang and Y.-C. Wu, "A novel rapid assessment of mental stress by using PPG signals based on deep learning," *IEEE Sensors J.*, vol. 22, no. 21, pp. 21232–21239, Nov. 2022, doi: 10.1109/JSEN.2022.3208427.
[15] M. Zubair and C. Yoon, "Multilevel mental stress detection using ultra-short pulse rate variability series," *Biomed. Signal Process. Control*, vol. 57, Mar. 2020, Art. no. 101736, doi: 10.1016/j.bspc.2019.101736.
[16] S. Heo, S. Kwon, and J. Lee, "Stress detection with single PPG sensor by orchestrating multiple denoising and peak-detecting methods," *IEEE Access*, vol. 9, pp. 47777–47785, 2021, doi: 10.1109/ACCESS.2021.3060441.
[17] B. Correia, N. Dias, P. Costa, and J. M. Pêgo, "Validation of a wireless Bluetooth photoplethysmography sensor used on the earlobe for monitoring heart rate variability features during a stress-inducing mental task in healthy individuals," *Sensors*, vol. 20, no. 14, p. 3905, Jul. 2020, doi: 10.3390/S20143905.
[18] P. Kalra and V. Sharma, "Mental stress assessment using PPG signal a deep neural network approach," *IETE J. Res.*, vol. 69, no. 2, pp. 879–885, Feb. 2023, doi: 10.1080/03772063.2020.1844068.
[19] N. Mukherjee, S. Mukhopadhyay, and R. Gupta, "Real-time mental stress detection technique using neural networks towards a wearable health monitor," *Meas. Sci. Technol.*, vol. 33, no. 4, Apr. 2022, Art. no. 044003, doi: 10.1088/1361-6501/ac3aae.
[20] H. Barki and W.-Y. Chung, "Mental stress detection using a wearable in-ear plethysmography," *Biosensors*, vol. 13, no. 3, p. 397, Mar. 2023, doi: 10.3390/bios13030397.
[21] J. Held, A. Vîslă, C. Wolfer, N. Messerli-Bürgy, and C. Flückiger, "Heart rate variability change during a stressful cognitive task in individuals with anxiety and control participants," *BMC Psychol.*, vol. 9, no. 1, 2021, Art. no. 44, doi: 10.1186/s40359-021-00551-4.
[22] C.-C. Wu, I.-W. Chen, and W.-C. Fang, "An implementation of motion artifacts elimination for PPG signal processing based on recursive least squares adaptive filter," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Oct. 2017, pp. 1–4, doi: 10.1109/BIOCAS.2017.8325141.
[23] S. S. Haykin and B. Van Veen. (2002). *Signals and Systems*. Accessed: May 18, 2023. [Online]. Available: https://www.wiley.com/en-us/Signals+and+Systems%2C+2nd+Edition-p-9780471164746
[24] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2017, pp. 5999–6009. Accessed: Jul. 19, 2023.
[25] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," Oct. 2020, *arXiv:2010.11929*. Accessed: May 22, 2023.

Hika Barki received the B.Sc. degree in biomedical engineering from Jimma University, Jimma, Ethiopia, in 2017, and the master's degree in systems and biomedical engineering from Cairo University, Giza, Egypt, in 2021. He is pursuing the Ph.D. degree in AI convergence with Pukyong National University, Busan, South Korea.

He is a highly dedicated and motivated individual with a profound passion for the convergence of Biomedical Engineering and Artificial Intelligence. His academic journey has equipped him with a solid foundation in both Biomedical Engineering and Systems Engineering. His deep interest in integrating these fields with AI has inspired. As a Ph.D. student, he actively engages in cutting-edge research in the domain of AI convergence, with a specific focus on areas such as wireless sensor networks, ubiquitous healthcare, biomedical signal analysis, and embedded systems. His ultimate goal is to leverage AI and other advanced technologies to transform healthcare delivery and enhance individuals' overall quality of life worldwide.

Lionel Nkenyereye received the B.S. degree in data communication from the High Institute of Technology, Bujumbura, Burundi, in 2005, and the M.S. and Ph.D. degrees in computer engineering from Dong-Eui University, Busan, South Korea, in 2016 and 2019, respectively.

He worked as a Postdoctoral Fellow with Sejong University, Seoul, South Korea. He is currently working as a Postdoctoral Fellow with Pukyong National University, Busan. Prior to joining Dong-Eui University, he served as an Information Technology Officer with Finbank Burundi, Bujumbura, from 2012 to 2014. He also worked as a Software Developer with Computer Applied Limited Burundi, Bujumbura, from November 2008 to July 2012. His research interests encompass a wide range of topics, including vehicle-to-everything communication, 5G and the Internet of Things (IoT), software-defined networks, networked embedded systems, edge computing, data communication, and vehicular networks.

Wan-Young Chung (Senior Member, IEEE) received the B.S. and M.S. degrees in electronic engineering from Kyungpook National University, Daegu, South Korea, in 1987 and 1989, respectively, and the Ph.D. degree in sensor engineering from Kyushu University, Fukuoka, Japan, in 1998.

He served as an Associate Professor with Semyung University, Jecheon, South Korea, from 1993 to 1999 and subsequently with Deongseo University, Busan, South Korea, from 1999 to 2008. Since September 2008, he has held the position of Full Professor with the Department of Electronic Engineering, Pukyong National University, Busan. His research interests include encompass a wide range of areas, including ubiquitous healthcare, wireless sensor network applications, and gas sensors.