# MuFuBP-Net: A Multimodal Fusion Network for Cuffless Blood Pressure Estimation Using Dual-Feature Pipeline With Probabilistic Feature Encoder

Farhad Hassan, *Graduate Student Member, IEEE*, Mubashir Ali, *Graduate Student Member, IEEE*, Zubair Akbar, *Graduate Student Member, IEEE*, Jingzhen Li, *Member, IEEE*, Yuhang Liu, Weihao Wang, Lixin Guo, and Zedong Nie, *Senior Member, IEEE*

***Abstract*—Cuffless blood pressure (BP) estimation is critical for managing growing concerns about hypertension and cardiovascular diseases. Despite recent advancements in multimodal (ECG and PPG) BP estimation methods, which have achieved varying degrees of success, several challenges remain to be addressed. These include capturing the full spectrum of BP-relevant information, redundant feature spaces, and handling the multigrade classification. To address these issues, we propose a Multimodal Fusion BP Network (MuFuBP-Net), featuring a novel dual-feature pipeline architecture designed to extract hierarchical and modality-specific features from both ECG and PPG signals. Additionally, the Cascading Cross-Feature Enhancer (CCFE) module integrates multiple fusion strategies with a squeeze-and-excitation mechanism to apply channel-wise attention to spatial features, enabling dynamic re-weighting. We also employed a Sequence Context Network (SCN) module to capture global sequential features. Subsequently, a Probabilistic Feature Encoder (PFE) encodes the multilevel features from both pipelines into a compact latent space, preserving their discriminative characteristics. Our approach achieved MAE $\pm$ SDE of 2.99 $\pm$ 4.37 mmHg (SBP) and 2.63 $\pm$ 4.19 mmHg (DBP) on MIMIC-II, and 2.27 $\pm$ 4.15 mmHg (SBP) and 1.63 $\pm$ 2.96 mmHg (DBP) on MIMIC-III dataset, meeting AAMI, BHS, and IEEE grade A standards. The proposed approach demonstrated competitive results compared to existing techniques, highlighting its significance as a reliable solution for cuffless BP monitoring.***

***Index Terms*—Cuffless blood pressure (BP), electrocardiogram (ECG), feature fusion, multimodal learning, photoplethysmography (PPG), probabilistic feature encoder (PFE), squeeze and excitation.**

## I. Introduction

HYPERTENSION is a prevalent cardiovascular condition that causes severe health complications [1]. It adversely affects various organs, leading to significant morbidity and mortality [2]. According to the World Health Organization (WHO), approximately one-third of the global adult population is affected by hypertension and timely diagnosis can prevent severe health consequences [3]. Clinical practice primarily employs two approaches for monitoring blood pressure (BP), which are direct (invasive) and intermittent (non-invasive) techniques [4]. The direct method enables continuous arterial blood pressure (ABP) measurement and involves cannulation of a major artery, which poses risks such as infection, thrombosis, and arterial damage [5]. On the other hand, intermittent monitoring typically employs inflatable cuffs and does not provide continuous BP data [6]. The accuracy of this method depends entirely on the observer's auditory acuity and the sensitivity of the stethoscope [7], [8]. Hence, recent studies have explored cuffless BP monitoring techniques using electrocardiography (ECG) and photoplethysmography (PPG) signals which are widely available in wearable devices. Among these, PPG-based methods for BP estimation present several significant advantages that make them an appealing approach for use in modern healthcare settings. These methods enhance patient comfort and reduce the risk of complications. Additionally, they are highly cost-effective, as they rely on relatively simple optical sensors that can be mass-produced at low cost, making them accessible for widespread use. This affordability, combined with seamless integration into wearable devices such as smartwatches and fitness

Received 2 January 2025; revised 7 April 2025; accepted 17 April 2025. Date of publication 23 April 2025; date of current version 7 October 2025. This work was supported in part by the Noncommunicable Chronic Diseases-National Science and Technology Major Project under Grant 2024ZD0532000, Grant 2024ZD0532002, in part by the National Natural Science Foundation of China under Grant 62173318, in part by the Science and Technology Service Network Plan of CAS-Huangpu Special Project under Grant STS-HP-202203, and in part by the Key Laboratory of Biomedical Imaging Science and System, Chinese Academy of Sciences. *(Corresponding author: Zedong Nie.)*

Farhad Hassan, Mubashir Ali, and Zubair Akbar are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China, and also with the University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: farhad@siat.ac.cn; mubashir@siat.ac.cn; zubair@siat.ac.cn).

Jingzhen Li, Yuhang Liu, and Zedong Nie are with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: lijz@siat.ac.cn; yh.liu2@siat.ac.cn; zd.nie@siat.ac.cn).

Weihao Wang and Lixin Guo are with the Department of Endocrinology, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing 100006, China (e-mail: wangweihaoedu@126.com; glx1218@163.com).

Digital Object Identifier 10.1109/JBHI.2025.3563852

trackers, facilitates convenient and continuous BP monitoring outside clinical settings, empowering individuals to track their health in real time [9], [10]. However, their performance is often affected by factors such as susceptibility to motion artifacts, variability in arterial stiffness, and the lack of a reliable temporal reference. Therefore, combining both modalities with deep learning approaches has demonstrated promising results due to the auto-extraction of high-fidelity features.

Baek et al. has utilized convolutional neural networks (CNNs) with dilated and strided convolutions to extract deep features from ECG and PPG through a shared feature matrix for BP prediction [11]. However, with local receptive fields and pooling layers, CNNs are limited in their ability to effectively address temporal variations. Su et al. incorporated Long Short-Term Memory (LSTM) networks to effectively capture temporal-scale features [12]. To integrate the strengths of both architectures, Kamanditya et al. proposed a hybrid CNN-LSTM method that generates a combined feature vector for spatial patterns and temporal information. This approach enhances the accuracy of BP prediction [13]. Li et al. incorporated bidirectional long short-term memory (BiLSTM), which can process information in both the forward and backward directions, thus capturing both past and future contexts. This resulted in a more comprehensive temporal feature representation, which further improved prediction accuracy [14]. Nevertheless, these methods encounter difficulties in maintaining relevant information over long sequences, which can cause the model to lose focus on important distant time steps. Therefore, Eom et al. incorporated attention mechanisms that selectively emphasize the most salient features of ECG and PPG signals, focusing on key temporal aspects across sequences. The feature vectors demonstrated better performance in BP estimation [15]. Koparır et al. introduced a methodology for transforming waveform data into two-dimensional images and extracting deep features using pre-trained CNN with transfer learning. They subsequently applied feature selection (recursive feature elimination algorithm) to identify 24 significant features of SBP and DBP prediction [16]. This approach was limited because static images restrict the capacity of the model to capture temporal dependencies fully. Wang et al. addressed the limitation of neglecting temporal dynamics in Koparır et al.'s work by employing a novel technique that transforms time-series data into images using visibility graph (VG) techniques. These VG were processed using pre-trained CNN models with only the final dense layer retrained on the BP dataset [17]. Tang et al. proposed a knowledge distillation transfer-learning approach utilizing a teacher-student (TS) model. ResNet-18 was combined with a gated recurrent unit (GRU) and a convolutional block attention module (CBAM) to capture both spatial and temporal features. The teacher model guided the student model to inherit deep features, thus improving the computational efficiency while maintaining predictive accuracy [18]. Similarly, in [19] SMART-BP two-stage framework model demonstrated the potential of transfer learning by pretraining on a high-quality dataset (Mindray) and fine-tuning on the MIMIC dataset. Considering the importance of establishing long-term correlations for feature sequences, Huang et al. proposed a method that integrates transformer encoders and stacked attention GRUs to extract high-level features from multi-source input, which enhances memory and recalibrates feature representations through attention mechanisms and achieves robust performance [20].

Although the achievements in the aforementioned multimodality-based studies are significant, still they have several limitations. Firstly, ECG and PPG exhibit higher complexity and nonlinear behavior, which prevents simple feature analysis from extracting all information contained in these signals [21], [22]. Therefore, advanced feature extraction methods are necessary to capture the full spectrum of BP-relevant information. Secondly, previous studies have consistently demonstrated superior accuracy in predicting diastolic blood pressure (DBP) over systolic blood pressure (SBP). This disparity can be partially attributed to the heightened variability of SBP as it exhibits greater sensitivity to rapid physiological alterations [23], [24]. Furthermore, an additional significant factor is the inherent imbalance in the multi-output framework, wherein favors one output over the other [25], [26]. This imbalance creates inconsistencies that are unsuitable for cuffless BP monitoring because the same method was classified into two different grades based on the key standards of the British Hypertension Society (BHS) [27] the Association for the Advancement of Medical Instrumentation (AAMI) [28]. Thus, an adaptive dynamic reweighting approach is necessary to ensure the fairness of the results. It adjusts the contributions of different features or channels based on their significance, ensuring that more informative or reliable features are given higher importance, whereas less relevant or noisy features contribute less to the final prediction. [29]. Thirdly, many studies have extracted features directly from ECG and PPG signals without effectively addressing the issue of feature redundancy. These studies relied on the network's inherent ability to optimize features. Notably, numerous researchers have shown that these signals inherently exhibit high levels of correlation due to physiological overlaps in cardiovascular dynamics, resulting in redundant information within the feature space [30], [31]. For that reason, it is essential to effectively exploit the correlation information between these signals to retain discriminative features for accurate BP estimation.

To overcome these challenges, we developed MuFuBP-Net, a multimodal fusion network using ECG and PPG to fuse hierarchical and modality-specific features with compact latent space features inspired by the VAE encoder, exploiting the correlation information between different physiological signals. The comprehensive framework is illustrated in Fig. 1, and the primary contributions are as follows:

1) We developed MuFuBP-Net with a dual-feature extraction pipeline that learns hierarchical and modality-specific dependencies for BP estimation. The MuFuBP-Net enhances discriminative feature extraction and handles the imbalance multitask regression problem between SBP and DBP.
2) We introduced the Cascading Cross-Feature Enhancer (CCFE) module, which integrates multimodal features, applies channel-wise attention to spatial features for dynamic re-weighting, and employs cross-modal learning to further enhance feature interactions.

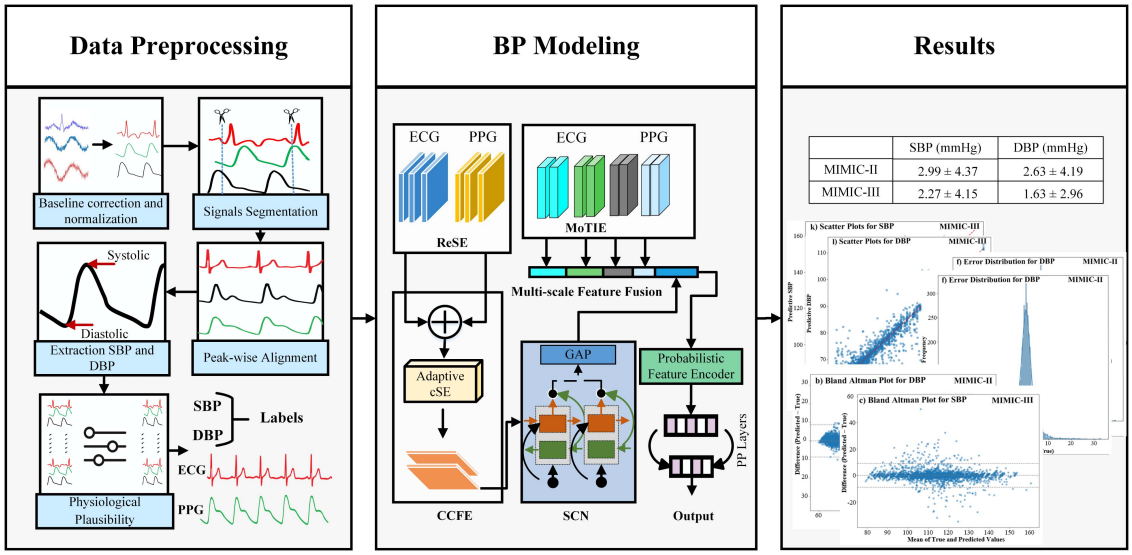


Fig. 1. Overall interactive framework for cuffless blood pressure (BP) monitoring using electrocardiogram (ECG) and photoplethysmogram (PPG) physiological signals.

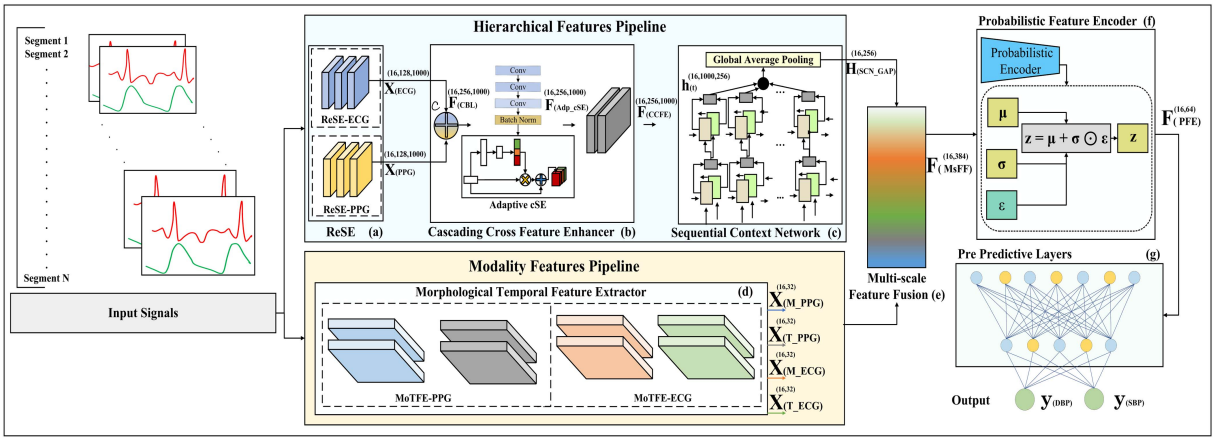


Fig. 2. Architecture of MuFuBP-NET including its modules (a) ReSE: Residual Self Encoding, (b) CCFE: Cascading-Cross Feature Enhancer, (c) SCN: Sequence Context Network, (d) MoTFE: Morphological and Temporal Feature Extractor, (e) Multi-scale Features Fusion, (f) Probabilistic Feature Encoder, and (e) Pre-Predictive Linear Layers.

3) We incorporated a Probabilistic Feature Encoder to enhance discriminative features by compact latent space with the reparameterization technique. The proposed approach demonstrates superior performance compared to state-of-the-art methods on MIMIC-II and -III datasets.

The following sections of paper are structured as follows: Section II provides a comprehensive overview of the proposed method, Section III describes the details of the experimental setup, and Section IV presents the results. Section V contains a discussion and limitations, and Section VI concludes the findings of this study.

## II. Methods

Fig. 2 illustrates the proposed MuFuBP-Net, a dual feature pipeline composed of seven interactive modules: ReSE, CCFE, SCN, MoTIE, MsFF, and PP Layers. Given a non-overlapped segment $S = \{S_{ecg}, S_{ppg}\} \in \mathbb{R}^{2\times L}$ with length $L$ as the input. The ReSE module comprises three residual blocks for each modality $S_{ecg}, S_{ppg}$ to focus on hierarchical features while maintaining signal fidelity $X_{PPG}, X_{ECG} \in \mathbb{R}^{128}$. These features were fed into the CCFE module, which consisted of three submodules: a cross-blend layer that fused the input features $F_{CBL} = CBL(X_{ECG}, X_{PPG}) \in \mathbb{R}^{256}$, adaptive $cSE$ that incorporates three convolution layers with channel-wise squeeze and excitation for feature refinement and recalibration $F_{AdpcSE} = AdpcSE(F_{CBL}) \in \mathbb{R}^{256}$, and cross-modal layer, thereby enhancing the capture of complementary information from signals $F_{CCFE} = CCFE(F_{AdpcSE}) \in \mathbb{R}^{256}$. Furthermore, to extract global sequential context-aware features applied SCN module $H_{SCN} = SCN(F_{CCFE}) \in \mathbb{R}^{256}$. The MoTIE module targeted localized aspects of the signals and extracted morphological $X_{M,ECG}, X_{M,PPG} \in \mathbb{R}^{32}$ and temporal features

$X_{T,ECG}, X_{T,PPG} \in \mathbb{R}^{32}$. The MsFF module combines hierarchical features with modality-specific features as: $F_{MsFF} = c_d(H_{SCN}, X_{M,ECG}, X_{M,PPG}, X_{T,ECG}, X_{T,PPG}) \in \mathbb{R}^{384}$. To reduce redundant features from combined features $F_{MsFF}$, incorporate the PFE Module and encode them into a compact latent space as: $F_{PFE} = PFE(F_{MsFF}) \in \mathbb{R}^{64}$. Finally, the PP layers module maps the $F_{PFE}$ into BP values $y_{SBP}, y_{DBP} \in \mathbb{R}$.

### A. Residual Self-Encoding (ReSE)

The Residual Self-Encoding (ReSE) comprises two parallel-stacked sub-blocks, ReSE-ECG and ReSE-PPG, which process the ECG and PPG signals independently but in a similar manner. Each block consists of three sequential residual blocks $R_1, R_2$, and $R_3$ that progressively increase the number of channels from 1 to 128. Each $R_i$ incorporates two convolutional layers $C_1$ and $C_2$, followed by BatchNorm $B$, and ReLU activation $\sigma$, represented by the network operation $N_i$. The linear transformation $L_i(X)$ is applied within $R_i$, where $i$ denotes the residual block number. Shortcut connections $S_i$ in each $R_i$ facilitate gradient flow, enable the learning of residual functions, and ensure that the network focuses on refining features without entirely altering the input. The output of each residual block $R_i$ as follows:

$$R_i(X) = N_i(L_i(X)) + S_i(X) \tag{1}$$

where $S_i$ is defined as follows:

$$S_i(X) = \begin{cases} X, & \text{if } D_{i-1} = D_i \\ C_{i,s}(X_i, \omega_{i,s}^N, b_{i,s}^N), & \text{if } D_{i-1} \neq D_i \end{cases} \tag{2}$$

Whenever the input dimension $D_{i-1}$ is differs from the output dimension $D_i$, a convolution operation $C_{i,s}$ is applied to $X_i$ using weights $\omega_{i,s}^N$ and biases $b_{i,s}^N$ to align the dimensions. The functions $F_{\text{PPG}}$ and $F_{\text{ECG}}$ are learned through the residual blocks $R_1, R_2$, and $R_3$ with learnable parameters $\Theta_{\text{PPG}}$ and $\Theta_{\text{ECG}}$, respectively. These functions transform the input signals $S_{\text{PPG}}$ and $S_{\text{ECG}}$ into the output feature maps $X_{\text{PPG}}$ and $X_{\text{ECG}}$ as follows:

$$X_{\text{PPG}} = F_{\text{PPG}}(S_{\text{PPG}}; \Theta_{\text{PPG}}) \in \mathbb{R}^{128} \tag{3}$$

$$X_{\text{ECG}} = F_{\text{ECG}}(S_{\text{ECG}}; \Theta_{\text{ECG}}) \in \mathbb{R}^{128} \tag{4}$$

### B. Cascading Cross Feature Enhancer (CCFE)

The Cross-Channel Feature Extraction (CCFE) module comprises three sub-modules: Cross-Blend Layer, Adaptive cSE, and Cross-Modal Layer. The Cross-Blend layer $F_{\text{CBL}}$ performs feature fusion to combine the $X_{\text{PPG}}$ and $X_{\text{ECG}}$ features of both modalities. The Adaptive cSE submodule is based on three convolutional layers: $C_1$ for channel mixing, $U_1 = C_1(F_{\text{CBL}})$; $C_2$, a depth-wise convolution for spatial feature extraction, $U_2 = C_2(U_1)$; and $C_3$, a point-wise convolution for feature combination, $U_3 = C_3(U_2)$. This is followed by a batch normalization $B$, such that $U_4 = B(U_3)$, after which a channel-wise squeeze-and-excitation mechanism is applied. This mechanism involves adaptive average pooling $A_{\text{avg}}$ to aggregate spatial information $z = A_{\text{avg}}(U_4)$ from each channel across the dimension. The pooled vector $z$ is passed through a linear layer $l_1$ and ReLU activation $\sigma$ with a reduction ratio of 16, $z_1 = \sigma(l_1(z))$. Subsequently, $z_1$ is fed into a second linear layer $l_2$ and a sigmoid activation function $F_{\text{sig}}$: $z_2 = F_{\text{sig}}(l_2(z_1))$. The vector $z_2$ is broadcast and applied to the original feature map $U_4$ via element-wise multiplication $\otimes$ for channel-wise recalibration $U_{\text{SE}} = U_4 \otimes z_2$. This channel-wise recalibration is designed to refine feature representations by selectively enhancing the most important channels while reducing the influence of less relevant ones. It begins with a pooling operation that compresses the input feature map into a compact representation, capturing essential global information [32]. Finally, element-wise addition $\oplus$ is performed between the result of this operation and the output of $F_{\text{CBL}}$, followed by ReLU activation $\sigma$ as follows:

$$F_{\text{AdpcSE}} = U_{\text{SE}} \oplus F_{\text{CBL}} \in \mathbb{R}^{256} \tag{5}$$

The Cross-Modal Layer submodule utilizes two residual blocks $R_1$ and $R_2$, each comprising two 1D convolutional layers $C_1, C_2$ with $3 \times 1$ kernels and padding, followed by BatchNorm $B$ and ReLU activation $\sigma$. A shortcut connection $S_i$ enables the input $F_{\text{AdpcSE}}$ to bypass these layers, facilitating gradient flow. The first residual block $R_1$ is learned through function $f_1$ with the input $F_{\text{AdpcSE}}$ as: $r_1 = \sigma(f_1(F_{\text{AdpcSE}}) + F_{\text{AdpcSE}})$. To further refine the feature representations learned in $R_1$, the second residual block $R_2$ applies function $f_2$ to the input $r_1$ as: $r_2 = \sigma(f_2(r_1) + r_1)$. The final output of the Cross-Channel Feature Extraction module $F_{\text{CCFE}}$ is expressed as follows:

$$\begin{aligned} F_{\text{CCFE}} = {} & \sigma(f_2(\sigma(f_1(F_{\text{AdpcSE}}) + F_{\text{AdpcSE}}))) \\ & + \sigma(f_1(F_{\text{AdpcSE}}) + F_{\text{AdpcSE}}) \in \mathbb{R}^{256} \end{aligned} \tag{6}$$

### C. Sequence Context Network (SCN)

To address long-range temporal dependencies across the sequence in both modalities, the Sequence Context Network module incorporates bidirectional sequence learning, from $t = 1$ to $t = T$ and $t = T$ to $t = 1$, with the concatenation of the forward hidden state $h_t^{\text{fwd}}$ and backward hidden state $h_t^{\text{bwd}}$. This module is based on two layer architecture of LSTM. The first LSTM layer $l_1$ processes the input sequence and produces an output $h_t^{(\text{bi},l_1)}$ for each time step $t$, where $h_t^{(\text{bi},l_1)} = [h_t^{\text{fwd}(1)}, h_t^{\text{bwd}(1)}]$. The second LSTM layer $l_2$ takes the output $h_t^{(\text{bi},l_1)}$ from layer $l_1$ as input and processes it to produce the final output $h_t^{(\text{bi},l_2)} = [h_t^{\text{fwd}(2)}, h_t^{\text{bwd}(2)}]$. Furthermore, global average pooling $G_{\text{avg}}$ is applied to summarize the features learned from the entire sequence into a single feature vector as follows:

$$H_{\text{SCN\_GAP}} = G_{\text{avg}}(h_t^{(\text{bi},l_2)}) = \frac{1}{T}\sum_{t=1}^{T} h_t^{(\text{bi},l_2)} \in \mathbb{R}^{256} \tag{7}$$

### D. Morphological and Temporal Feature Extractor (MoTFE)

The Morphological and Temporal Feature Extractor (MoTFE) module comprises two sub-module MOTFE-ECG and MOTFE-PPG. These modules capture the local morphological and temporal features from the input raw signals $S_{\text{ECG}}$ and $S_{\text{PPG}}$, respectively. In both MOTIE-ECG and MOTIE-PPG, the Morphological Feature Extractor (MFE) block utilizes a convolutional layer

$C$ followed by ReLU activation $\sigma$ and adaptive average pooling $A_{\text{avg}}$. The convolution filters $\omega_{M_{\text{ECG}}}$ and $\omega_{M_{\text{PPG}}}$ are applied to the $S_{\text{ECG}}$ and $S_{\text{PPG}}$ signals to compute the feature as follows:

$$m_{\text{ECG}} = S_{\text{ECG}} * \omega_{M_{\text{ECG}}} \tag{8}$$

$$m_{\text{PPG}} = S_{\text{PPG}} * \omega_{M_{\text{PPG}}} \tag{9}$$

Added bias terms $b_{M_{\text{ECG}}}$ and $b_{M_{\text{PPG}}}$ to the convolution outputs $X_{M_{\text{ECG}}}$ and $X_{M_{\text{PPG}}}$ for the $S_{\text{ECG}}$ and $S_{\text{PPG}}$ signals, to extract local shape-based characteristics as follows:

$$X_{M_{\text{ECG}}} = A_{\text{avg}}\left(\sigma\left(m_{\text{ECG}} + b_{M_{\text{ECG}}}\right)\right) \in \mathbb{R}^{32} \tag{10}$$

$$X_{M_{\text{PPG}}} = A_{\text{avg}}\left(\sigma\left(m_{\text{PPG}} + b_{M_{\text{PPG}}}\right)\right) \in \mathbb{R}^{32} \tag{11}$$

In the Temporal Feature Extractor (TFE) block, a 1D convolutional layer $C$ with ReLU activation $\sigma$ and adaptive max pooling $M_{\text{adapt}}$. The input signals $S_{\text{ECG}}$ and $S_{\text{PPG}}$ underwent convolution with filters $\tau_{T_{\text{ECG}}}$ and $\tau_{T_{\text{PPG}}}$, respectively, to derive temporal features as follows:

$$T_{\text{ECG}} = S_{\text{ECG}} * \tau_{T_{\text{ECG}}} \tag{12}$$

$$T_{\text{PPG}} = S_{\text{PPG}} * \tau_{T_{\text{PPG}}} \tag{13}$$

To incorporate bias terms $b_{T_{\text{ECG}}}$ and $b_{T_{\text{PPG}}}$ with convolution outputs $X_{T_{\text{ECG}}}$ and $X_{T_{\text{PPG}}}$ for the input signals $S_{\text{ECG}}$ and $S_{\text{PPG}}$, respectively, in order to facilitate the identification of temporal patterns, as follows:

$$X_{T_{\text{ECG}}} = M_{\text{adapt}}\left(\sigma\left(T_{\text{ECG}} + b_{T_{\text{ECG}}}\right)\right) \in \mathbb{R}^{32} \tag{14}$$

$$X_{T_{\text{PPG}}} = M_{\text{adapt}}\left(\sigma\left(T_{\text{PPG}} + b_{T_{\text{PPG}}}\right)\right) \in \mathbb{R}^{32} \tag{15}$$

### E. Multi-Scale Feature Fusion (MsFF)

In this multiscale feature fusion module, hierarchical features $H_{\text{SCN\_GAP}}$ and modality-specific features $X_{M_{\text{ECG}}}, X_{M_{\text{PPG}}}, X_{T_{\text{ECG}}}, X_{T_{\text{PPG}}}$ are concatenated into a single vector, which is mathematically expressed as follows:

$$F_{\text{MsFF}} = c_d\left[H_{\text{SCN\_GAP}}, X_{M_{\text{ECG}}}, X_{M_{\text{PPG}}}, X_{T_{\text{ECG}}}, X_{T_{\text{PPG}}}\right] \in \mathbb{R}^{384} \tag{16}$$

Where $c$ denotes concatenation along feature dimension $d$.

### F. Probabilistic Feature Encoder (PFE)

The Probabilistic Feature Encoder (PFE) module, inspired by the encoder component of Variational Autoencoders (VAE) [33], is designed to reduce redundant features and enhance the discriminative power of the extracted features. The mean vector $\mu(F_{\text{MsFF}})$ is computed as $\mu(F_{\text{MsFF}}) = \omega_\mu\,\phi(\omega_1 F_{\text{MsFF}} + b_1) + b_\mu$, and the logarithm of the variance as $\log\sigma^2(F_{\text{MsFF}}) = \omega_{\log\sigma^2}\,\phi(\omega_1 F_{\text{MsFF}} + b_1) + b_{\log\sigma^2}$. Here, $\phi$ represents the ReLU activation function, and $\omega_1$, $\omega_\mu$, and $\omega_{\log\sigma^2}$ are weight matrices, while $b_1$, $b_\mu$, and $b_{\log\sigma^2}$ are bias vectors of the respective layers. To enable efficient backpropagation through the stochastic layer, the reparameterization trick $\exp(0.5 \times \log\sigma^2(F_{\text{MsFF}}))$ is applied. This trick samples $\epsilon$ from a standard normal distribution $N(0, I)$, is used to scale and shift the mean and variance to sample the latent variable $F_{\text{PFE}}$, where $\odot$ denotes element-wise multiplication. We performed systematic experimentation to determine the optimal features for the PFE. Specifically, we used 16, 32, 64, 128, and 256 features for PFE. Based on performance, we selected 64 features for the final PFE block. Therefore, the combined 384-dimensional feature vector $F_{\text{MsFF}}$ is mapped to a 64-dimensional latent space $F_{\text{PFE}}$ as follows:

$$F_{\text{PFE}} = \mu(F_{\text{MsFF}}) + \exp\left(\frac{1}{2}\log\sigma^2(F_{\text{MsFF}})\right) \odot \epsilon \tag{17}$$

### G. Pre Predictive Layers (PP Layers)

The Pre-Predictive Layers module, also known as the regressor, comprises two fully connected layers. The first layer, $\text{fc}_1$, transforms the 64-dimensional input into 32 dimensions, while $\text{fc}_2$ reduces it further from 32 to 16. Both $\text{fc}_1$ and $\text{fc}_2$ are separated by ReLU activation $\sigma$. This process can be mathematically expressed as follows:

$$y_2 = \sigma\left(\omega_2\left(\sigma\left(\omega_1 \cdot F_{\text{MsFF}} + b_1\right)\right) + b_2\right) \in \mathbb{R}^{16} \tag{18}$$

Where $\omega_1$ and $\omega_2$ represent the weights, and $b_1$ and $b_2$ denote the biases.For the final prediction, two separate outputs are used to estimate SBP and DBP, with weight matrices $\omega_{\text{sbp}}$ and $\omega_{\text{dbp}}$ (each of size $16 \times 1$) and scalar biases $b_{\text{sbp}}$ and $b_{\text{dbp}}$, respectively as follows:

$$y_{\text{SBP}} = \omega_{\text{sbp}} \cdot y_2 + b_{\text{sbp}} \tag{19}$$

$$y_{\text{DBP}} = \omega_{\text{dbp}} \cdot y_2 + b_{\text{dbp}} \tag{20}$$

## III. Experimental Setup

### A. Dataset

This study used two subsets from the Multi-Parameter Intelligent Monitoring in Intensive Care (MIMIC) II and III. The MIMIC Waveform Database includes physiological data from ICU patients including ECG, PPG, and ABP signals, sampled at 125Hz. A subset of MIMIC-II, compiled by Kachuee et al. [34], includes data from 942 subjects, while a subset of MIMIC-III, curated by Samsung R&D Institute Brazil et al. [35], contains data from 1524 subjects.

### B. Data Preparation

The data preparation pipeline comprises several components, as illustrated in Fig. 3 to address various types of anomalies, noise, baseline drift, and abnormal waveforms.

*1) Baseline Correction and Normalization:* To mitigate noise, reduce artifacts, and correct the baseline-wander from both modality signals, a Butterworth bandpass filter with PPG (0.5–20Hz) and ECG (2–20Hz) was employed. After that, the signals were normalized to a 0–1 range to standardize the signal amplitudes. ABP remained unaltered to preserve critical BP information, as modifications could potentially impact systolic and diastolic blood pressure derivations inappropriately.

*2) Signals Segmentation:* In some previous studies, overlapped segmentation was used, which led to data leakage [36], [37], [38]. To mitigate this issue, this study adopted a non-overlapping segmentation approach that ensures no portion of the data is shared across segments. We also assessed segments

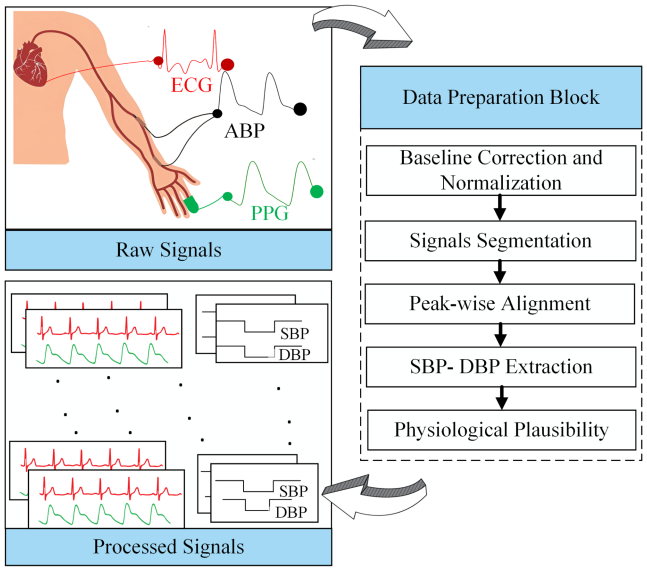


Fig. 3. Data preparation block.

of 4, 8, and 12 seconds. The 4-second segments exhibited high variability and reduced accuracy, whereas the 12-second segments imposed computational burden without yielding significant performance improvements. The 8-second segments effectively balance the capture of essential signal dynamics with computational efficiency, rendering them particularly suitable for real-time applications. The MIMIC-II and MIMIC-III datasets were sampled at a rate of 125Hz, resulting in 1,000 data points per segment ($125 \times 8 = $ t1,000). This segmentation approach ensured a consistent representation of the ECG, PPG, and corresponding ABP, thereby facilitating the modeling of physiological patterns. Any segment shorter than the specified length was discarded to ensure that only valid 8-second segments were retained. The MIMIC-III subset was labeled for 30 seconds; therefore, we omitted the last 6 seconds of data to maintain consistency across both MIMIC-II and MIMIC-III datasets.

*3) Peak-Wise Alignment:* Peak synchronization and cardiovascular cycle demarcation across ECG, PPG, and ABP modalities were achieved by identifying characteristic peaks within 8-second segments using BioSPPy library algorithms [39], followed by synchronizing PPG and ABP waveforms relative to ECG R-peaks through temporal offset computation and circular phase shifts, effectively ensuring temporal coherence across all these modalities.

*4) Extraction of SBP/DBP:* SBP and DBP were derived from the ABP signal using a peak detection method named `find_peaks` in SciPy library [40]. For the height and distance parameters, an adaptive threshold method is employed for the height and distance parameters, which is widely adopted in biomedical signal processing to enhance robustness. This approach utilizes statistical thresholds to reduce false positives in noisy physiological signals. The height threshold is set to $\mu_{\text{ABP}} + \sigma_{\text{ABP}}$ and the distance threshold calculated as $D(t)$, where $f_s$ is the frequency rate of the sample. The mean and

TABLE I
HYPERPARAMETER SEARCH SPACE AND OPTIMAL SELECTION

| Hyperparameter | Pool of Values | Selected Value |
|---|---|---|
| Optimizer | {SGD, Adam, RMSprop} | Adam |
| Learning Rate | {1e-2, 1e-3, 1e-4, 1e-5} | 1e-4 |
| Weight Decay | {0, 1e-6, 1e-5, 1e-4} | 1e-5 |
| Batch Size | {8, 16, 32, 64} | 16 |
| ReduceLROnPlateau Factor | {0.1, 0.2, 0.5} | 0.1 |
| ReduceLROnPlateau Patience | {3, 5, 7, 10} | 5 |

standard deviation of the ABP signal are computed as follows:

$$\mu_{\text{ABP}} = \frac{1}{N}\sum_{i=1}^{N} \text{ABP}_i \tag{21}$$

$$\sigma_{\text{ABP}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\text{ABP}_i - \mu_{\text{ABP}})^2} \tag{22}$$

The distance threshold is calculated as follows:

$$D(t) = \frac{(0.6 \times f_s) + (1.0 \times f_s)}{2} \tag{23}$$

*5) Physiological Plausibility:* To ensure physiological plausibility, performed four key operations. First, a BP range filter was applied to eliminate implausible SBP and DBP values, constraining them within clinically ranges (80–180mmHg for SBP and 60–130mmHg for DBP). Second, hemodynamic parameter validation was conducted to ensure that the pulse pressure (the difference between SBP and DBP) was maintained between 20 and 60mmHg and that SBP consistently exceeded DBP, adhering to established cardiovascular principles. Third, statistical outlier detection was performed to remove extreme values by utilizing quantile calculations to determine the 0.1th and 99.9th percentiles for the signals, establishing range filters that eliminated likely errors while preserving natural variability. Finally, each 8-second segment was required to contain a minimum of six systolic peaks, ensuring sufficient data density and consistency across segments, corresponding to a minimum heart rate of 45bpm. Following the preprocessing phase, both datasets contained 40,012 and 22,138 valid 8-second segments from MIMIC-II and MIMIC-III, respectively.

## C. Implementation Settings

All experiments were performed on an RTX A6000 GPU within a software environment comprising Python 3.8, PyTorch 1.7.1, CUDA 11.0, and cuDNN 8.0.5. We employed the 80:10:10 ratio with the Hold-Out Out-of-sample (HOO) strategy to split the dataset. The optimal hyperparameters for the proposed model were selected through an extensive grid search strategy, which systematically explored combinations of hyperparameters to identify the best configuration for model performance as shown in Table I. The final configuration included the Adam optimizer for expeditious convergence with minimal adjustments, a learning rate of 1e-4 to achieve equilibrium between speed and precision, and a weight decay of 1e-5 to regularize the model and

TABLE II
THE ARCHITECTURAL SETUP AND RESULTS OF DIFFERENT ABLATION STUDIES

| Ablation Study Structure and Description | MIMIC-II | | | | MIMIC-III | | | |
|---|---|---|---|---|---|---|---|---|
| | SBP | | DBP | | SBP | | DBP | |
| | MAE | SDE | MAE | SDE | MAE | SDE | MAE | SDE |
| Case 1: ReSE +Fusion + SCN | 4.31 | 5.81 | 3.85 | 4.79 | 3.19 | 5.15 | 2.69 | 4.02 |
| Case 2: ReSE+SCN+MoTFE+Fusion | 4.23 | 5.44 | 3.62 | 4.65 | 3.08 | 4.93 | 2.34 | 3.97 |
| Case 3: ReSE+CCFE+SCN+MoTFE+Fusion | 3.75 | 4.87 | 3.05 | 4.49 | 2.65 | 4.74 | 2.09 | 3.45 |
| **Case 4: ReSE+CCFE+SCN+MoTFE+Fusion+PFE** | **2.99** | **4.37** | **2.63** | **4.19** | **2.27** | **4.15** | **1.63** | **2.96** |

ReSE: Residual Self Encoding; CCFE: Cascading-Cross Feature Enhancer; SCN: Sequence Context Network; MoTFE: Morphological and Temporal Feature Extractor; PFE: Probabilistic Feature Encoder;SBP: Systolic Blood Pressure; DBP: Diastolic Blood Pressure; MEA: Mean Absolute Error; SDE: Standard Deviation of Error

mitigate overfitting. Early stopping was implemented to monitor the validation loss and terminate the training on the performance plateau. The ReduceLROnPlateau scheduler dynamically modulated the learning rate based on the validation performance, with a reduction factor of 0.1 and patience of five epochs. A batch size of 16 was selected to accommodate the variability in the physiological signals, facilitating more granular updates to the model weights. These hyperparameters were determined to optimize the performance, expedite convergence, and minimize overfitting, thereby enhancing the efficiency of the MuFuBP-Net framework.

### D. Performance Evaluation Metrics

The performance of MuFuBP-Net was evaluated using established metrics in accordance with the international standards and protocols set by the AAMI [28], BHS [27], and IEEE [41] for BP estimation. These standards stipulate that key metrics include the Mean Error (ME), Mean Absolute Error (MAE), and Standard Deviation Error (SDE). The mathematical expressions for these metrics are as follows:

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}\left(y_{\mathrm{pred}_i} - y_{\mathrm{true}_i}\right) \tag{24}$$

$$\mathrm{SDE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(e_i - \mathrm{ME}\right)^2} \tag{25}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_{\mathrm{pred}_i} - y_{\mathrm{true}_i}\right| \tag{26}$$

Moreover, scatter, error distribution, Bland-Altman, and actual vs predict plots serve as valuable tools for evaluating the concordance between the model's predictions and reference values.

## IV. RESULTS

### A. Ablation Studies

We conducted several ablation experiments on MIMIC-II and MIMIC-III datasets to assess the impact of different modules of the proposed model as presented in Table II.

Case 1 employed the ReSE module with fusion and sequence context networks as the baseline model of our ablation experiments. Case 2 extended the baseline architecture by incorporating the MoTFE module. It outperformed Case 1 and confirmed that hierarchical and modality-specific dependencies fusion is a more favorable approach for BP estimation. Moreover, in Case 3, the performance further improved because of the CCFE module, which incorporated multilevel fusion with a dynamic weight feature representation. Notably, the significant performance was enhanced by incorporating the VAE-inspired Probabilistic Feature Encoder (PFE) module to exploit the correlation information. Hence, these ablation experiments showed that the proposed approach facilitates the extraction of discriminative features, thereby enhancing the model's performance in BP estimation.

TABLE III
INTERNATIONAL STANDARDS AND PROTOCOLS FOR BLOOD PRESSURE ESTIMATION

| Standard / Protocol | Metrics | MIMIC-II | | MIMIC-III | |
|---|---|---|---|---|---|
| | | SBP | DBP | SBP | DBP |
| IEEE | MAE (mmHg) | 2.99 | 2.63 | 2.27 | 1.63 |
| | Grade | A | A | A | A |
| AAMI | ME (mmHg) | -0.03 | -0.03 | -0.01 | -0.03 |
| | SDE (mmHg) | 4.37 | 4.19 | 4.15 | 2.96 |
| | Grade | Pass | Pass | Pass | Pass |
| BHS | $\leq$ 5 mmHg (%) | 82.39 | 86.77 | 88.94 | 93.90 |
| | $\leq$ 10 mmHg (%) | 96.14 | 96.49 | 96.32 | 98.45 |
| | $\leq$ 15 mmHg (%) | 98.91 | 98.73 | 98.45 | 99.38 |
| | Grade | A | A | A | A |

### B. BP Estimation Performance

MuFuBP-Net achieved MAEs of 2.99 $\pm$ 4.37 mmHg (SBP) and 2.63 $\pm$ 4.19 mmHg (DBP) with MEs of $-0.03$ (SBP) and 0.03 (DBP) on the MIMIC-II dataset, and MAEs of 2.27 $\pm$ 4.15 mmHg (SBP) and 1.63 $\pm$ 2.96 mmHg (DBP) with MEs of 0.14 (SBP) and minus-0.06 (DBP) on the MIMIC-III dataset, meeting all key standard criteria with grade A for blood pressure estimation established by AAMI, BHS, and IEEE as shown in Table III.

Fig. 4 presents a comprehensive performance evaluation of the proposed model on both MIMIC-II and MIMIC-III datasets: (a)–(d) present Bland-Altman plots for SBP and DBP, assessing the agreement between the predicted and actual BP values. These plots indicate a strong level of consistency, with the majority of the data points falling within the 95(% ) confidence interval ($\pm$ 1.96 SDE). The differences between predicted and actual BP values are centered around zero, signifying minimal systematic bias. Notably, the MIMIC-III plots (c and d) exhibit a slightly narrower spread compared to MIMIC-II (a and b), indicating

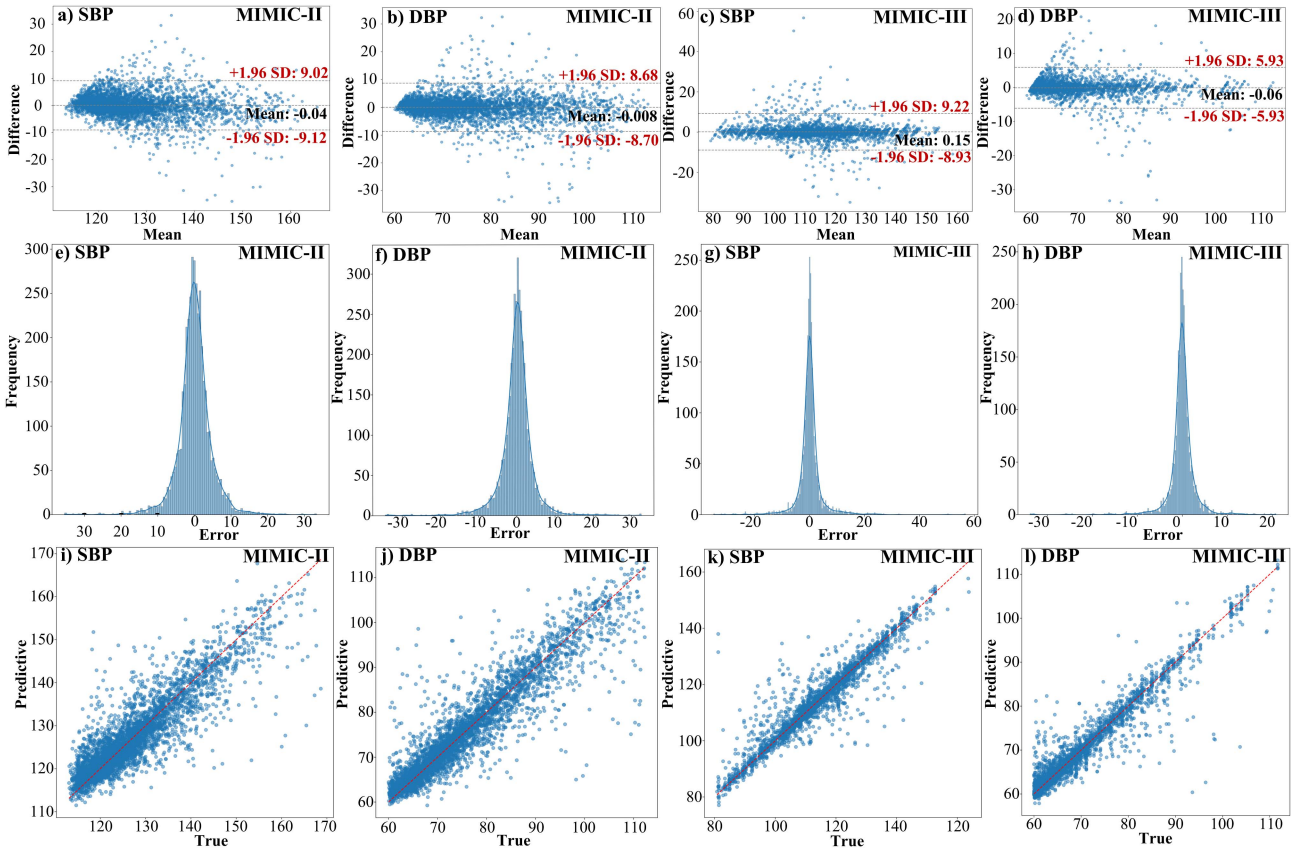


Fig. 4. Comprehensive Performance Evaluation Plots: (a)–(d) Bland-Altman, (e)–(h) Error distribution, and (i)–(l) Scatter plots for SBP and DBP on MIMIC-II and MIMIC-III, respectively.

reduced variability and enhanced consistency in model performance on the MIMIC-III dataset. Sub-figures (e)– (h) show the error distribution histograms for the SBP and DBP. These histograms show a narrow, peaked distribution centered around 0 mmHg, reflecting minimal error and suggesting that most predictions lie within a small error range, with only minor deviations from the actual values. The peak at zero indicates that the model is unbiased and neither underestimates nor overestimates the BP values. The MIMIC-III histograms (sub-figures g and h) revealed a slightly narrower distribution than those of MIMIC-II (sub-figures e and f), further highlighting that the model achieved higher accuracy and reduced variability in predictions on the MIMIC-III dataset. Finally, scatter plots for SBP and DBP (sub-figures i– l) show a strong linear correlation, with data points closely clustered around the regression line, indicating high agreement. Clustering appears tighter in the MIMIC-III plots (sub-figures k and l), highlighting that the model predictions are closer to actual values.

Fig. 5 presents a segment-wise comparison between the actual and predicted blood pressure (BP) values for the MIMIC-II (a) and MIMIC-III (b) datasets. Each segment corresponds to 8 seconds of physiological data, resulting in a total of 75 segments, which collectively represent 10 minutes of continuous BP monitoring. In the plots, red lines represent SBP, and blue lines represent DBP. Solid lines indicate the ground truth values, whereas dashed lines denote the model's predictions. The close alignment between the actual and predicted curves clearly demonstrates the model's performance.

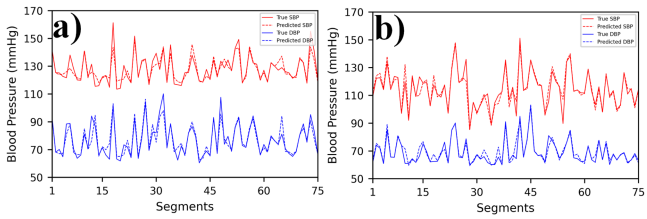


Fig. 5. Segment-wise comparison of actual and predicted SBP and DBP values for the a) MIMIC-II and b) MIMIC-III datasets.

Table IV presents the cross-validation results of the proposed model using the K-fold method, wherein both datasets were partitioned into K=10 subsets and evaluated across different folds to ensure the validity of the proposed approach.

### C. Explainability Analysis Using LIME

We applied the Local Interpretable Model-Agnostic Explanations approach to interpret the decision-making process of our MuFuBP-Net, which integrates both ECG and PPG features for BP estimation. The feature contribution plots for the MIMIC-II and MIMIC-III datasets are shown in Figs. 6 and 7, respectively. These plots reveal a consistent pattern across both datasets:

TABLE IV
TEN-FOLD CROSS-VALIDATION RESULTS OF PROPOSED APPROACH

| Fold | MIMIC-III | | | | MIMIC-II | | | |
|---|---|---|---|---|---|---|---|---|
| | SBP | | DBP | | SBP | | DBP | |
| | MAE | SDE | MAE | SDE | MAE | SDE | MAE | SDE |
| 1 | 2.36 | 4.48 | 1.66 | 2.87 | 3.04 | 4.43 | 2.78 | 4.39 |
| 2 | 2.25 | 4.22 | 1.57 | 2.93 | 2.84 | 4.19 | 2.45 | 3.95 |
| 3 | 2.16 | 4.12 | 1.54 | 2.86 | 3.03 | 4.45 | 2.69 | 4.26 |
| 4 | 2.16 | 3.98 | 1.54 | 2.91 | 2.96 | 4.37 | 2.58 | 4.12 |
| 5 | 2.21 | 4.85 | 1.56 | 2.78 | 2.99 | 4.35 | 2.57 | 4.22 |
| 6 | 2.33 | 4.12 | 1.66 | 2.96 | 3.10 | 4.52 | 2.78 | 4.42 |
| 7 | 2.21 | 4.05 | 1.57 | 3.02 | 2.95 | 4.25 | 2.57 | 4.19 |
| 8 | 2.36 | 4.11 | 1.73 | 2.90 | 2.97 | 4.28 | 2.63 | 4.15 |
| 9 | 2.37 | 4.28 | 1.71 | 3.21 | 3.06 | 4.43 | 2.67 | 4.23 |
| 10 | 2.33 | 4.32 | 1.75 | 3.15 | 2.97 | 4.38 | 2.58 | 3.94 |
| Avg. | 2.27 | 4.15 | 1.63 | 2.96 | 2.99 | 4.37 | 2.63 | 4.19 |

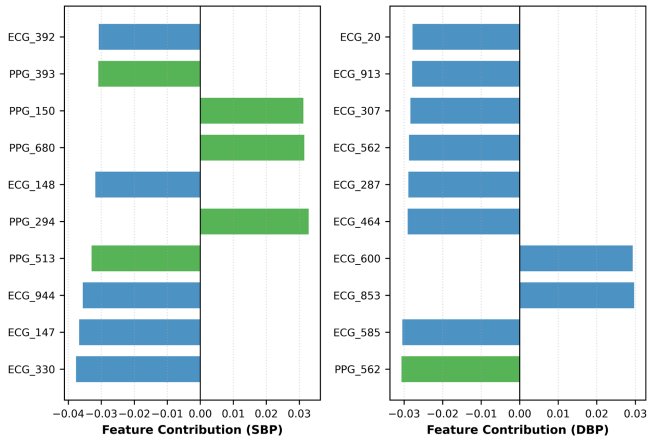


Fig. 6. LIME-based feature contribution analysis for explainability of MuFuBP-Net on the MIMIC-II.

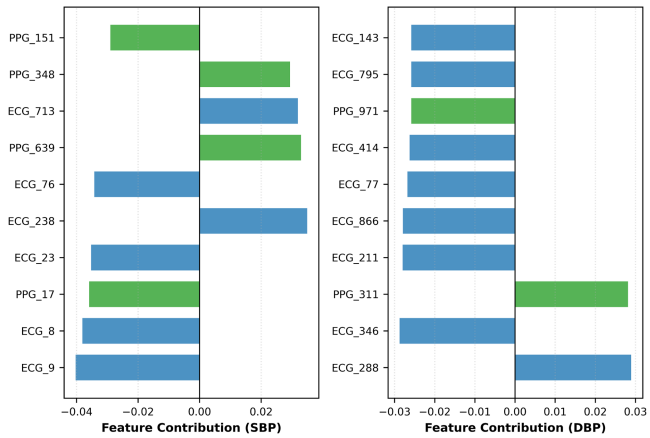


Fig. 7. LIME-based feature contribution analysis for explainability of MuFuBP-Net on the MIMIC-III.

ECG-derived features have a greater influence on the model predictions, whereas PPG features have a lesser contribution. This finding aligns with the established clinical knowledge that ECG signals have a stronger correlation with BP variations than PPG signals. From a physiological perspective, ECG signals offer a more direct representation of cardiovascular dynamics and autonomic regulation, both of which are closely associated with BP changes. On the other hand, PPG provides valuable hemodynamic information but is more susceptible to external factors, such as motion artifacts and variations in peripheral perfusion, which may weaken its correlation with BP fluctuations.

TABLE V
RESULT COMPARISON WITH DIFFERENT METHODS

| Dataset | Ref. | Findings | | | |
|---|---|---|---|---|---|
| | | SBP | | DBP | |
| | | MAE | SDE | MAE | SDE |
| MIMIC-II | [11] | 5.32 | 5.54 | 3.38 | 3.42 |
| | [14] | 4.63 | 14.5 | 3.15 | 6.44 |
| | [42] | 8.02 | 7.06 | 4.54 | 4.84 |
| | [43] | 4.18 | 5.87 | 2.47 | **3.52** |
| | [17] | 6.17 | 8.46 | 3.66 | 5.36 |
| | [44] | 4.70 | 7.60 | **2.01** | 3.90 |
| | [16] | 5.63 | 8.83 | 2.82 | 5.60 |
| | **MuFuBP-NET** | **2.99** | **4.37** | 2.63 | 4.19 |
| MIMIC-III | [45] | 4.41 | 6.11 | 2.91 | 4.23 |
| | [46] | 3.67 | 4.81 | 2.79 | 3.62 |
| | [47] | 4.84 | 5.05 | 2.95 | 3.31 |
| | [13] | 5.30 | 7.24 | 3.29 | 4.76 |
| | [18] | 2.94 | 5.61 | 2.02 | 3.60 |
| | [20] | 3.91 | 5.65 | 2.29 | 3.01 |
| | **MuFuBP-NET** | **2.27** | **4.15** | **1.63** | **2.96** |

### D. Comparison With State-of-the-Art Studies

Table V presents a comparative analysis of different BP estimation methods across the MIMIC-II and MIMIC-III datasets. On MIMIC-II dataset, MuFuBP-Net achieved the lowest MEA (2.99) with SDE (4.37) for SBP, outperforming the existing methods and indicating a high level of predictive accuracy and stability. In DBP, our model also achieved strong results with an MEA of 2.63 with an SDE of 4.19, positioning it among the most reliable methods for DBP prediction. Although there exists a study on the MIMIC-II dataset with a lower MEA with SDE for DBP than MuFuBP-Net. Their MEA with SDE for SBP is considerably higher. This discrepancy suggests that their model may be more tailored to SBP predictions, but lacks a balance in handling DBP accurately, which is critical in hypertensive care. On MIMIC-III dataset, MuFuBP-Net demonstrates superior performance across both SBP and DBP, with MEA values of 2.27 for SBP and 1.63 for DBP, and SDE values of 4.15 and 2.96, respectively.

In addition, our proposed approach also outperformed several state-of-the-art lightweight methods on both the MIMIC-II and MIMIC-III datasets. On the MIMIC-II dataset, compared to BP-Net [48], which reported MAEs of 5.16 mmHg for SBP and 2.89 mmHg for DBP, our method achieved MAEs that are 42.1% lower for SBP and 9.0% lower for DBP. On the MIMIC-III dataset, our method demonstrated superior performance compared to multiple lightweight approaches. Specifically, it achieved MAEs that are 35.5% lower for SBP and 25.9% lower for DBP compared to the transfer learning approach [10] (MAEs: 3.52 mmHg SBP, 2.20 mmHg DBP); 32.0% lower for SBP and 24.6% lower for DBP compared to the Lightweight Diffusion Model (DiffCNBP) [49] (MAEs: 3.34 mmHg SBP, 1.87 mmHg DBP); and 49.9% lower for SBP and 51.6% lower for DBP

compared to the multimodal lightweight CNN-LSTM [50] (MAEs: 4.53 mmHg SBP, 3.37 mmHg DBP).

## V. Discussion

BP monitoring has significant potential for improving individual health outcomes, as uncontrolled BP can lead to severe complications. This study introduces a cuffless BP monitoring framework that utilizes ECG and PPG physiological signals, incorporating an 8-second non-overlapping window size with at least six systolic peaks. For a normal heart rate range (60–100 beats per minute), these six systolic peaks in 8-seconds accurately represent cardiac activity. Thus, this 8-second window size optimally captures meaningful variability in the signals, which is crucial for distinguishing between different BP levels. The shared feature extraction pipeline struggles to fully capture the complementary nature of ECG and PPG signals due to the potential dilution of their distinct physiological characteristics when processed through a single shared pathway. This limitation may result in the loss of critical information required for accurate BP estimation. Whereas, the dual-feature pipeline enhances the feature space by offering a more comprehensive representation of BP-related dynamics, which is essential for precise BP estimation. To the best of our knowledge, this is the first method that utilizes such a dual-feature pipeline with a VAE-inspired probabilistic feature encoder architecture for BP estimation. The hierarchical pipeline employs stacked residual layers separately for ECG and PPG signals, and these layers capture low- and high-level features and shortcut connections in residual blocks to prevent signal degradation with depth. It maintains the critical physiological characteristics of both modalities when they pass through layers. A key component of this approach is the Cross-Channel Feature Extractor (CCFE) block, which combines the characteristics of each modality. Following this, depthwise separable convolutions, which integrate depthwise convolutions for spatial feature extraction and pointwise convolutions for channel mixing when combined with a squeeze-and-excitation (SE) mechanism, enable efficient feature extraction with reduced computational complexity. The SE mechanism incorporates channel-wise attention to emphasize salient features by explicitly modeling interchannel dependencies. This allows the network to adaptively recalibrate feature responses, enhancing the most informative features while suppressing the less relevant ones. While traditional attention mechanisms primarily focus on spatial dimensions without explicit channel modeling, the SE method provides a structured and computationally efficient approach to capture complex interchannel relationships. This results in an improved feature representation and enhanced overall model performance with minimal additional computational overhead. In addition, a Cross-Modal Layer performs an intermediate fusion of recalibrated features at a deeper level using residual learning techniques to identify and integrate complex dependencies between the ECG and PPG data. BP is influenced not only by the immediate cardiac cycle but also by its patterns over time. Therefore, an SCN block was employed to capture sequential dependencies by recognizing BP patterns across multiple cardiac cycles. This approach enabled the model to identify both forward and backward signal patterns, facilitating a comprehensive understanding that is essential for continuous BP monitoring. To evaluate the effectiveness of the SCN block, a transformer-based architecture, specifically a Linformer, was implemented. However, it struggles to capture fine-grained temporal variations in ECG and PPG signals, which are critical for the accurate estimation of BP. Parallel to this, the modality-specific pipeline extracts signal-specific characteristics from each modality. It captures local shape-based and short-term temporal characteristics, focusing on immediate signal structure. This enables the extraction of recurring, rapid patterns relevant to BP fluctuations and enhances the model's understanding of BP-related dynamics. The outputs from both pipelines are then fused to create a rich combined feature representation. These aggregated features are subsequently reduced to a lower-dimensional latent space by the VAE-inspired probabilistic feature encoder, which effectively eliminates redundant information while preserving discriminative characteristics. Our proposed method addresses the traditional challenges associated with multimodal BP estimation approaches. Compared with existing methods, our proposed method demonstrates significant improvements in MAE and SDE across both the MIMIC-II and MIMIC-III datasets, outperforming state-of-the-art models. The integration of hierarchical and modality-specific features, along with the probabilistic feature encoder, was the key factor contributing to the enhanced performance of our model. Notably, the MIMIC-III dataset generally yielded lower prediction errors than MIMIC-II, likely owing to its improved data quality and completeness, as supported by previous studies. Our proposed method meets the key standards established by international organizations for hypertension management, including the AAMI, BHS, and IEEE. However, it is important to note that the SBP predictions exhibited a higher MAE than the DBP predictions. This discrepancy may be attributed to the intrinsic characteristics of SBP, such as cardiac output and vascular resistance [23], [24]. To address this, we incorporated an adaptive channel-wise Squeeze-and-Excitation (cSE) mechanism that dynamically reweights features to emphasize the most physiologically relevant patterns. This approach significantly contributes to reducing the discrepancy between SBP and DBP prediction accuracy. MuFuBP-Net has a parameter size of 2.00 million, a memory footprint of 7.62 MB (FP32), and a computational complexity of 1.108 GFLOPs. These resource requirements make the model feasible for implementation in clinical environments, such as ICUs, where devices with moderate computational resources are typically available. Furthermore, cloud-based implementation offers another practical use case for the model and enables real-time data processing with higher predictive performance.

Despite the significant results of this study, cuffless BP measurement has limitations that persist in this domain. Firstly, optimizing the model for deployment on resource-constrained wearable devices remains challenging, as computational efficiency is a bottleneck despite the use of quantization techniques. Secondly, while the model demonstrates strong performance on the clinically accepted MIMIC-II and MIMIC-III waveform datasets, these datasets primarily represent ICU patients, limiting immediate applicability to broader real-world settings. In

the future, we aim to extend this approach to wearable applications by applying transfer learning and knowledge distillation techniques. A teacher-student framework could enable a complex teacher model trained on multiple datasets to enhance the model's ability to generalize across different populations while guiding a lightweight student model and improving adaptability to diverse physiological conditions.

## VI. Conclusion

This study presents MuFuBP-Net, a novel multimodal fusion architecture for noninvasive BP estimation utilizing ECG and PPG signals. The core strength of MuFuBP-Net lies in its innovative dual-feature extraction pipeline, which is hierarchical and modality-specific, coupled with a probabilistic feature encoder. This dual-feature extraction framework comprises multiple modules to capture a comprehensive representation of features, including channel-wise attention for dynamic reweighting. Furthermore, these features were fused, and redundant information was eliminated through a probabilistic feature encoder module motivated by a variational autoencoder (VAE). The proposed approach enhanced the discriminative power of the extracted features. Experimental results on publicly available datasets demonstrate that the model outperforms current state-of-the-art methods for cuffless BP monitoring. It also addresses the imbalance problem in the simultaneous estimation of systolic and diastolic blood pressures. These findings underscore the potential of MuFuBP-Net as a reliable solution for real-world cuffless BP monitoring, thereby establishing a robust foundation for advancements in intelligent healthcare systems.

## References


[1] K. T. Mills, A. Stefanescu, and J. He, "The global epidemiology of hypertension," *Nature Rev. Nephrol.*, vol. 16, no. 4, pp. 223–237, 2020.

[2] B. Zhou, P. Perel, G. A. Mensah, and M. Ezzati, "Global epidemiology, health burden and effective interventions for elevated blood pressure and hypertension," *Nature Rev. Cardiol.*, vol. 18, no. 11, pp. 785–802, 2021.

[3] W. H. Organization, *Global Report on Hypertension: The Race Against a Silent Killer*. Geneva, Switzerland: World Health Organization, 2023.

[4] M. Cepeda, P. Pham, and D. Shimbo, "Status of ambulatory blood pressure monitoring and home blood pressure monitoring for the diagnosis and management of hypertension in the US: An up-to-date review," *Hypertension Res.*, vol. 46, no. 3, pp. 620–629, 2023.

[5] S. Mathew, M. Archana, and R. Sharma, "A comparative study of upper limb and lower limb blood pressure measured by auscultatory and oscillometric method with intra-arterial blood pressure in hemodynamically unstable patients," *Dent. Med. Res.*, vol. 10, no. 2, pp. 44–48, 2022.

[6] M. Wijnberge, B. V. D. Ster, A. P. Vlaar, M. W. Hollmann, B. F. Geerts, and D. P. Veelo, "The effect of intermittent versus continuous non-invasive blood pressure monitoring on the detection of intraoperative hypotension, a sub-study," *J. Clin. Med.*, vol. 11, no. 14, 2022, Art. no. 4083.

[7] B. G. Celler, J. Basilakis, K. Goozee, and E. Ambikairajah, "Non-invasive measurement of blood pressure-why we should look at BP traces rather than listen to Korotkoff sounds," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, IEEE, 2015, pp. 5964–5967.

[8] B. G. Celler, P. Le, J. Basilakis, and E. Ambikairajah, "Improving the quality and accuracy of non-invasive blood pressure measurement by visual inspection and automated signal processing of the Korotkoff sounds," *Physiol. Meas.*, vol. 38, no. 6, pp. 1006–1022, 2017.

[9] D. H. Nguyen et al., "Predicting blood pressures for pregnant women by PPG and personalized deep learning," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 1, pp. 5–16, Jan. 2025.

[10] J. Leitner, P.-H. Chiang, and S. Dey, "Personalized blood pressure estimation using photoplethysmography: A transfer learning approach," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 1, pp. 218–228, Jan. 2022.

[11] S. Baek, J. Jang, and S. Yoon, "End-to-end blood pressure prediction via fully convolutional networks," *IEEE Access*, vol. 7, pp. 185458–185468, 2019.

[12] P. Su, X.-R. Ding, Y.-T. Zhang, J. Liu, F. Miao, and N. Zhao, "Long-term blood pressure prediction with deep recurrent neural networks," in *Proc. 2018 IEEE EMBS Int. Conf. Biomed. Health Inform.*, IEEE, 2018, pp. 323–328.

[13] B. Kamanditya, Y. N. Fuadah, N. Q. Mahardika T, and K. M. Lim, "Continuous blood pressure prediction system using Conv-LSTM network on hybrid latent features of photoplethysmogram (PPG) and electrocardiogram (ECG) signals," *Sci. Rep.*, vol. 14, no. 1, 2024, Art. no. 16450.

[14] Y.-H. Li, L. N. Harfiya, K. Purwandari, and Y.-D. Lin, "Real-time cuffless continuous blood pressure estimation using deep learning model," *Sensors*, vol. 20, no. 19, 2020, Art. no. 5606.

[15] H. Eom et al., "End-to-end deep learning architecture for continuous blood pressure estimation using attention mechanism," *Sensors*, vol. 20, no. 8, 2020, Art. no. 2338.

[16] H. M. Koparir and Ö. Arslan, "Cuffless blood pressure estimation from photoplethysmography using deep convolutional neural network and transfer learning," *Biomed. Signal Process. Control*, vol. 93, 2024, Art. no. 106194.

[17] W. Wang, P. Mohseni, K. L. Kilgore, and L. Najafizadeh, "Cuff-less blood pressure estimation from photoplethysmography via visibility graph and transfer learning," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 5, pp. 2075–2085, May 2022.

[18] H. Tang et al., "Blood pressure estimation based on PPG and ECG signals using knowledge distillation," *Cardiovasc. Eng. Technol.*, vol. 15, no. 1, pp. 39–51, 2024.

[19] C. Ma et al., "SMART-BP: SEM-ResNet and auto-regressor based on a two-stage framework for noninvasive blood pressure measurement," *IEEE Trans. Instrum. Meas.*, vol. 73, 2024, Art. no. 2503718.

[20] Z. Huang, J. Shao, P. Zhou, B. Liu, J. Zhu, and D. Fang, "Continuous blood pressure monitoring based on transformer encoders and stacked attention gated recurrent units," *Biomed. Signal Process. Control*, vol. 99, 2025, Art. no. 106860.

[21] E. H. Houssein, M. Kilany, and A. E. Hassanien, "ECG signals classification: A review," *Int. J. Intell. Eng. Inform.*, vol. 5, no. 4, pp. 376–396, 2017.

[22] M. A. Motin, C. K. Karmakar, and M. Palaniswami, "Selection of empirical mode decomposition techniques for extracting breathing rate from PPG," *IEEE Signal Process. Lett.*, vol. 26, no. 4, pp. 592–596, Apr. 2019.

[23] Y. Zhou, Z. Tan, Y. Liu, and H. Cheng, "Fully convolutional neural network and PPG signal for arterial blood pressure waveform estimation," *Physiol. Meas.*, vol. 44, no. 7, 2023, Art. no. 075007.

[24] S. Yang, W. S. W. Zaki, S. P. Morgan, S.-Y. Cho, R. Correia, and Y. Zhang, "Blood pressure estimation with complexity features from electrocardiogram and photoplethysmogram signals," *Opt. Quantum Electron.*, vol. 52, pp. 1–16, 2020.

[25] H. Borchani, G. Varando, C. Bielza, and P. Larranaga, "A survey on multi-output regression," *Wiley Interdiscipl. Rev.: Data Mining Knowl. Discov.*, vol. 5, no. 5, pp. 216–233, 2015.

[26] Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi, "Delving into deep imbalanced regression," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 11842–11851.

[27] N. Atkins, "The british hypertension society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems," *J. Hypertens*, vol. 8, pp. 607–619, 1990.

[28] W. B. White et al., "National standard for measurement of resting and ambulatory blood pressures with automated sphygmomanometers," *Hypertension*, vol. 21, no. 4, pp. 504–509, 1993.

[29] C. Li, S. Ding, N. Zou, X. Hu, X. Jiang, and K. Zhang, "Multi-task learning with dynamic re-weighting to achieve fairness in healthcare predictive modeling," *J. Biomed. Inform.*, vol. 143, 2023, Art. no. 104399.

[30] H. Wang et al., "Non-invasive continuous blood pressure prediction based on ECG and PPG fusion map," *Med. Eng. Phys.*, vol. 119, 2023, Art. no. 104037.

[31] E. Finnegan, S. Davidson, M. Harford, P. Watkinson, L. Tarassenko, and M. Villarroel, "Features from the photoplethysmogram and the electrocardiogram for estimating changes in blood pressure," *Sci. Rep.*, vol. 13, no. 1, 2023, Art. no. 986.

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

[33] D. P. Kingma et al., "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[34] M. Kachuee, M. M. Kiani, H. Mohammadzade, and M. Shabany, "Cuffless blood pressure estimation algorithms for continuous health-care monitoring," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 4, pp. 859–869, Apr. 2017.
[35] V. V. Sanches et al., "MIMIC-BP: A curated dataset for blood pressure estimation," *Sci. Data*, vol. 11, no. 1, Art. no. 1233, 2024.
[36] T. Athaya and S. Choi, "An estimation method of continuous non-invasive arterial blood pressure waveform using photoplethysmography: A U-net architecture-based approach," *Sensors*, vol. 21, no. 5, 2021, Art. no. 1867.
[37] M. A. Mehrabadi, S. A. H. Aqajari, A. H. A. Zargari, N. Dutt, and A. M. Rahmani, "Novel blood pressure waveform reconstruction from photoplethysmography using cycle generative adversarial networks," in *Proc. 44th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, IEEE, 2022, pp. 1906–1909.
[38] W. Long and X. Wang, "BPNet: A multi-modal fusion neural network for blood pressure estimation using ECG and PPG," *Biomed. Signal Process. Control*, vol. 86, 2023, Art. no. 105287.
[39] P. Bota, R. Silva, C. Carreiras, A. Fred, and H. P. da Silva, "BioSPPy: A python toolbox for physiological signal processing," *SoftwareX*, vol. 26, 2024, Art. no. 101712.
[40] P. Virtanen et al., "Scipy 1.0: Fundamental algorithms for scientific computing in python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.
[41] I. S. Association et al., *IEEE Standard for Wearable Cuffless Blood Pressure Measuring Devices*, IEEE Standard 1708-2014, 2014.
[42] L.-P. Yao and Z.-l. Pan, "Cuff-less blood pressure estimation from photoplethysmography signal and electrocardiogram," *Phys. Eng. Sci. Med.*, vol. 44, no. 2, pp. 397–408, 2021.
[43] B. Huang, W. Chen, C.-L. Lin, C.-F. Juang, and J. Wang, "MLP-BP: A novel framework for cuffless blood pressure measurement with PPG and ECG signals based on MLP-mixer neural networks," *Biomed. Signal Process. Control*, vol. 73, 2022, Art. no. 103404.
[44] G. Bossavi, R. Yan, and M. Irfan, "A novel convolutional neural network deep learning implementation for cuffless heart rate and blood pressure estimation," *Appl. Sci.*, vol. 13, no. 22, 2023, Art. no. 12403.
[45] S. Baker, W. Xiang, and I. Atkinson, "A hybrid neural network for continuous and non-invasive estimation of blood pressure from raw electrocardiogram and photoplethysmogram waveforms," *Comput. Methods Programs Biomed.*, vol. 207, 2021, Art. no. 106191.
[46] G. Ma, L. Zheng, W. Zhu, X. Xing, L. Wang, and Y. Yu, "Prediction of arterial blood pressure waveforms based on multi-task learning," *Biomed. Signal Process. Control*, vol. 92, 2024, Art. no. 106070.
[47] W. Long, J. Li, and X. Wang, "Blood pressure estimation neural network using large kernel convolutional attention," in *Proc. 15th Int. Conf. Digit. Image Process.*, 2023, pp. 1–6.
[48] K. R. Vardhan, S. Vedanth, G. Poojah, K. Abhishek, M. N. Kumar, and V. Vijayaraghavan, "BP-Net: Efficient deep learning for continuous arterial blood pressure estimation using photoplethysmogram," in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl.*, 2021, pp. 1495–1500.
[49] C. Ma, L. Guo, H. Zhang, Z. Liu, and G. Zhang, "DiffCNBP: Lightweight diffusion model for IoMT-based continuous cuffless blood pressure waveform monitoring using PPG," *IEEE Internet Things J.*, vol. 12, no. 1, pp. 61–80, Jan. 2025.
[50] S. Baker, W. Xiang, and I. Atkinson, "A computationally efficient CNN-LSTM neural network for estimation of blood pressure from features of electrocardiogram and photoplethysmogram waveforms," *Knowl.-Based Syst.*, vol. 250, 2022, Art. no. 109151.