*BioMedInformatics*

MDPI

*Article*

# Hybrid Neural Network Models to Estimate Vital Signs from Facial Videos

**Yufeng Zheng**

University of Mississippi Medical Center, Jackson, MS 39216, USA; yzheng@umc.edu

**Abstract: Introduction**: Remote health monitoring plays a crucial role in telehealth services and the effective management of patients, which can be enhanced by vital sign prediction from facial videos. Facial videos are easily captured through various imaging devices like phone cameras, webcams, or surveillance systems. **Methods**: This study introduces a hybrid deep learning model aimed at estimating heart rate (HR), blood oxygen saturation level (SpO2), and blood pressure (BP) from facial videos. The hybrid model integrates convolutional neural network (CNN), convolutional long short-term memory (convLSTM), and video vision transformer (ViViT) architectures to ensure comprehensive analysis. Given the temporal variability of HR and BP, emphasis is placed on temporal resolution during feature extraction. The CNN processes video frames one by one while convLSTM and ViViT handle sequences of frames. These high-resolution temporal features are fused to predict HR, BP, and SpO2, capturing their dynamic variations effectively. **Results**: The dataset encompasses 891 subjects of diverse races and ages, and preprocessing includes facial detection and data normalization. Experimental results demonstrate high accuracies in predicting HR, SpO2, and BP using the proposed hybrid models. **Discussion**: Facial images can be easily captured using smartphones, which offers an economical and convenient solution for vital sign monitoring, particularly beneficial for elderly individuals or during outbreaks of contagious diseases like COVID-19. The proposed models were only validated on one dataset. However, the dataset (size, representation, diversity, balance, and processing) plays an important role in any data-driven models including ours. **Conclusions**: Through experiments, we observed the hybrid model's efficacy in predicting vital signs such as HR, SpO2, SBP, and DBP, along with demographic variables like sex and age. There is potential for extending the hybrid model to estimate additional vital signs such as body temperature and respiration rate.

**Keywords:** vital sign; facial video; convolutional neural network (CNN); convolutional long short-term memory (convLSTM); video vision transformer (ViViT); deep learning; telehealth



Academic Editor: Alexandre G. De Brevern

Received: 14 November 2024
Revised: 9 January 2025
Accepted: 14 January 2025
Published: 22 January 2025

**Citation:** Zheng, Y. Hybrid Neural Network Models to Estimate Vital Signs from Facial Videos. *BioMedInformatics* **2025**, *5*, 6. https://doi.org/10.3390/biomedinformatics5010006



## 1. Introduction

### *1.1. Background Introduction*

The primary vital signs, encompassing body temperature, heart rate (HR), respiration rate (RR), blood pressure (BP), and blood oxygen saturation (SpO2), play pivotal roles in identifying physiological decline, prompting additional scrutiny or intervention when necessary. Vigilant monitoring of these vital signs is especially critical in acute care settings, where patients are vulnerable to increased risks and require intensified surveillance. While conventional contact-based monitoring devices are widely used and generally reliable, they may be cumbersome or intrusive for patients. Non-contact, video-based methods present

a versatile and information-rich alternative that could facilitate novel improvements in patient care management.

Photoplethysmography (PPG) has emerged as a notable player in this field in the past decade due to its cost-effectiveness and non-intrusive nature, enabling the detection of subtle changes in reflected light caused by physiological activities. Remote PPG (rPPG), a contactless variant, has become a practical and widely available solution, leveraging cameras embedded in everyday devices such as smartphones or laptops. Since the inception of the initial rPPG-based approach, researchers have made significant advancements in evaluating cardiac functions, monitoring RR [1], gauging BP [2], and determining SpO2 [3]. Facial regions, being relatively unobstructed, are preferred for PPG signal extraction owing to their visibility. Nonetheless, challenges persist in accurately estimating vital signs through rPPG. Signal strength during video capture may be weak and prone to interference, while subject movement, including head motions, changes in posture, and facial gestures, remains a constant concern. Additionally, variations in indoor lighting and digital camera noise further compound interference issues. Furthermore, the absence of suitable datasets presents significant obstacles, particularly for deep learning approaches seeking end-to-end solutions.

Traditional analytical techniques such as independent component analysis (ICA), principal component analysis (PCA) [4,5], and wavelet analysis [1] encounter difficulties in handling unstable conditions like motion, changes in facial expressions, and variations in environmental lighting. In contrast, deep learning, which employs neural networks to directly predict vital signs, offers a promising alternative. Chen [6] proposed a convolutional attention network, integrating attention mechanisms into video-based vital sign assessment. This network automatically identifies and prioritizes regions of interest (ROIs) in videos, enabling accurate measurement of HR and RR, even in scenarios involving significant head movement.

To address the challenge of redundant spatial information in facial video data, Hu [7] introduced a model for extracting spatiotemporal facial features using short segments. They also introduced a Spatial-Temporal Attention Module to mitigate the effects of head motion and improve the reliability of HR estimation. In tackling noise from facial expressions across various facial regions, Lokendra [8] proposed a novel denoising-rPPG method for HR estimation. This method utilized a Temporal Convolution Network and integrated action units to filter noise from the temporal signal, thereby enhancing HR estimation accuracy. Yin [9] presented an HR estimation model incorporating an attention mechanism module to focus on the skin area, suppressing background noise, along with a spatiotemporal convolution module to improve estimation accuracy across different environments. Li [10] introduced an HR estimation approach based on a multi-level convolutional neural network comprising stages such as low-level face feature generation, 3D spatiotemporal stack convolution, multi-hierarchical feature fusion, and signal prediction.

Recent advancements in deep learning have significantly improved BP measurement via video analysis. Luo [11] devised a method for BP estimation utilizing smartphone cameras to capture facial videos with a focus on the subject's forehead. This approach employed Transdermal Optical Imaging for facial region selection and feature extraction. Subsequently, a Multilayer Perceptron was trained on 30 features obtained post-PCA dimensionality reduction to construct a BP prediction model. The Joint Approximation Diagonalization of Eigen-matrices (JADE) algorithm separated blind source signals from the green channel in the ROI, followed by PPG signal denoising and extraction of systolic and diastolic pressure peaks. Wu [12] introduced an end-to-end BP estimation model leveraging multichannel remote PPG input, incorporating a Generative Adversarial Network (GAN) called infoGAN to enhance data augmentation. Iuchi [13] explored a method for remotely

estimating continuous BP by utilizing pulse waves and spatial information. This involved transforming spatial pulse wave signals into spatial signals for each ROI contour. A ResNet was then trained on these contours to estimate continuous BP.

### 1.2. Related Work

Deep neural networks (DNNs) [14] and skin reflection models [15] have emerged as promising avenues to enhance remote physiological measurement. Methods grounded in skin reflection models explore diverse techniques to translate images from the RGB space to alternative color spaces, aiming to refine rPPG estimation. For instance, Haan et al. [15] proposed projecting images into the chrominance space to mitigate motion artifacts and improve rPPG estimation accuracy. In recent years, DNN approaches have shown remarkable efficacy in remote physiological measurement [14,16,17]. Hu et al. [18] developed a time-domain attention network to extract and fuse temporal information from multiple video segments for rPPG estimation. Song et al. [14] utilized a conditional GAN to translate chrominance signals from [15] into precise rPPG waveforms. Yu et al. [16] addressed rPPG estimation from highly compressed and low-resolution facial videos by integrating video enhancement networks, while Yu et al. [17] employed transformer blocks to capture the relationship among video frame features.

Despite their significant achievements, these methodologies require substantial amounts of facial videos paired with synchronously recorded PPG signals for effective model training. However, curating such annotated datasets poses challenges. Gideon et al. [19] tackled this issue by training the DNN in a self-supervised manner using unlabeled data: they resampled the input video to generate its negative counterpart and applied a contrastive loss between rPPG signals extracted from the input and resampled videos.

Hsu et al. [20] proposed a method for heart rate (HR) estimation from facial videos, where a 2D Time-Frequency Representation extracted from facial frames, captured over short time intervals, served as the input image (feature map) for a CNN model. Omer et al. [21] introduced a beat-to-beat blood pressure (BP) estimation system using facial videos, employing Transfer Learning based on pre-trained image deep learning networks. The 1D beat was transformed into a 2D image, and cleaning metrics were applied to both the training data and the extracted remote rPPG signal. Only the selected beats were utilized to train the DL network (ResNet-101) to predict BP. The mean absolute error (MAE) for estimated BP from real videos was 7.05 for systolic BP (SBP) and 5.62 for diastolic BP (DBP).

More recently, Cheng et al. [22] introduced a novel deep learning approach for remote blood oxygen saturation (SpO2) measurement using facial videos captured by consumer-level RGB cameras. The proposed models utilize spatial-temporal representations of facial video data, which are fed into CNNs to predict SpO2 levels. The best-performing model achieves a root mean squared error (RMSE) of 1.71%, surpassing the international standard of 4% for approved pulse oximeters.

Jaiswal et al. [23] proposed a video-based noise-resistant cardiopulmonary measurement method for contactless heart rate (HR) estimation, addressing challenges like weak physiological signals and noise caused by head movements, illumination variations, and device inconsistencies. A novel motion representation, derived using wavelet decomposition from subsequent frames, is projected to a CNN (ResNet-18) for HR estimation under diverse lighting and continuous motion. Experimental results on four benchmark datasets (e.g., RMSE = 7.21) demonstrate the method's superior performance for HR estimation under challenging conditions.

Lin et al. [24] introduced a dual-path estimation method for continuous and contactless monitoring of vital signs using remote photoplethysmography. The method addresses challenges like weak signals, body movements, and limited data generalization by combining video magnification and deep learning. The approach begins with automated facial region of interest (ROI) detection, tracking, and magnification. Features such as heart rate, pulse transit time (PTT), and pulse wave waveform characteristics are extracted from the steady wave components of the ROIs. A small CNN then estimates blood pressure from these features. The results are RMSE (HR, SBP, and DBP) = 5.33, 4.11, and 3.75, demonstrating the method's potential for future clinical applications.

Park et al. [25] proposed to measure facial respiratory rate (FRR) using facial videos and robust respiratory signals across varied environments. The method incorporates bounding box stabilization to remove high-frequency noise from face detection and real-time coordinate correction for accurate signal calculation. An ROI is defined on the face, and RGB data is transformed into the YCgCo color space. Respiratory signals are extracted using partial zero-padding-based Fast Fourier Transform (FFT) and inverse FFT (iFFT) applied to the Cg channel. The FRR is then determined by analyzing peak-to-peak intervals within the derived respiratory signal. Results (RMSE = 1.01) demonstrate the accuracy and reliability of the proposed approach, highlighting its potential for robust, noninvasive respiratory monitoring in diverse real-world conditions.

Zheng [26] introduced a method utilizing facial images and a CNN-SVM model to forecast body temperature. Facial images were captured using both a digital camera and a smartphone, and ResNet-50 was employed to extract facial features. Subsequently, an SVM model was trained to predict body temperature. The ResNet-50-SVM model yielded a low error rate (<0.35 °C) while maintaining a rapid processing speed. This contactless approach to body temperature measurement and monitoring proved particularly valuable during the COVID-19 pandemic period (2020–2023).

As summarized in Table 1, previous studies in the literature have estimated remote rPPG signals from facial videos and subsequently trained deep learning models to predict one single vital sign, typically heart rate or blood pressure. One significant problem is the lack of a compressive approach to estimate all vital signs from facial videos. In this study, we present a hybrid end-to-end model comprising CNN, convLSTM, and ViViT, which is capable of directly inferring multiple vital signs from facial videos. Our contributions are to develop one integrated neural network model for predicting all vital signs and to train the model on a fairly large dataset. The remainder of this paper is structured as follows: Section 2 outlines the preprocessing steps for the Vital Video Dataset. Section 3 discusses the deep learning neural network models employed. Section 4 presents the experimental findings. Finally, Section 5 provides a summary of the paper.

**Table 1.** Summary of the most recent studies for vital sign estimation from facial videos.

| Ref.\Study | Method | Vital Sign | Dataset (Size) | Result (RMSE) |
|---|---|---|---|---|
| [22] | CNN | SpO2 | VIPL-HR (107 subjects) | 1.71% |
| [23] | CNN (ResNet-18) | HR | VIPL-HR (107 subjects) | 7.21 |
| [24] | video magnification, CNN | HR, SBP, DBP | Private (288 videos) | 5.33, 4.11, 3.75 |
| [25] | FFT | RR | Private (unknown size) | 1.01 |
| [26] | CNN-SVM | Body temperature | Private (144 subjects) | 0.35 |
| (This Study) | Hybrid CNN model | HR, SpO2, SBP, DBP | VVD-Large (891 subjects) [27] | 5.55, 0.87, 4.07, 2.21 |

## 2. Dataset Preprocessing

The proposed NN model for vital sign estimation requires preprocessing, including normalization, face detection, and standardization.

### *2.1. Image Normalization and Face Detection*

Image normalization (intensity scaling/stretching) is defined as follows:

$$\mathbf{I}_N = (\mathbf{I}_0 - I_{Min})\frac{L_{Max} - L_{Min}}{I_{Max} - I_{Min}} + L_{Min} \quad (1)$$

where $\mathbf{I}_N$ is the normalized image, $\mathbf{I}_0$ is the original image, $I_{Min}$ and $I_{Max}$ are the minimum and maximum pixel values in $\mathbf{I}_0$, and $L_{Min}$ and $L_{Max}$ are the desired minimum and maximum pixel values in $\mathbf{I}_N$. For example, we may select $L_{Min}$ = 0 and $L_{Max}$ = 255.

Face detection is a critical task in image processing, serving as a pivotal step for subsequent operations. Various techniques have been developed for detecting faces within a single image, including knowledge-based methods, feature-invariant facial approaches, template-matching methods, and appearance-based methods. A comprehensive survey on face detection is available in [28]. Many methods leverage color information, such as identifying regions (skin maps) with similar skin color, effectively narrowing down the search area. One of the most successful algorithms for visual image processing is the Viola–Jones algorithm [29], introduced in 2001. Face detection methods efficiently identify multiple faces within an image, irrespective of their sizes and backgrounds, by leveraging spatial changes. The Viola–Jones (VJ) algorithm has become widely adopted for object detection, particularly in face detection. Viola and Jones developed this algorithm as a machine learning approach for object detection, prioritizing fast and accurate detection rates. The VJ method incorporates three key aspects. Firstly, it utilizes *integral images* [29,30] as an image representation structure, where features are computed by summing pixel values within rectangular areas, extending a method introduced by Papageorgiou et al. [31].

Although Viola and Jones acknowledge the simplicity of using rectangles, they enable efficient computational calculations [29,30], which aligns well with the Adaboost algorithm employed for learning features in the image. This extension of the Adaboost algorithm enables the system to select a set of features and train classifiers, a concept initially discussed by Freund and Schapire [32]. This learning mechanism enables the system to discern between integral representations of faces and background regions. Another significant aspect of the VJ method is the use of *cascading classifiers*. At each stage, the classifier either rejects instances (represented as sliding windows from a test image) based on a given feature value or forwards them for further processing. Initially, a large number of negative examples are swiftly eliminated. To expedite the system, Viola and Jones discard areas highly unlikely to contain the object. The image undergoes testing with the first cascade, with rejected areas no longer processed. Regions likely to contain the object are further assessed until all classifiers are evaluated, with areas in the final classifier deemed likely to contain the object (face).

### *2.2. Facial Image Standardization*

We apply general standardization to all facial images:

$$\mathbf{I}_S = \left(\mathbf{I}'_N - \mu\right)/\sigma \quad (2)$$

where $\mathbf{I}_S$ is the standardized image, $\mathbf{I}'_N$ is the cropped facial image after normalization using Equation (1) and face detection, and $\mu$ and $\sigma$ denote the mean and standard deviation of all faces ($\mathbf{I}'_N$ images) after normalization, respectively.

Every face was identified and isolated from every frame, then resized to 288 × 256 pixels, standardized, and normalized (with intensity stretched). The normalized facial videos are segmented into 30-frame clips as input data, while the outcome variables (outputs) consist of four vital signs (HR, SpO2, systolic BP, and diastolic BP), as well as two demographic variables: sex and age. These six outcome variables are treated as continuous variables, allowing for the training of a neural network (NN) model for regression. The predicted sex values can be thresholded to generate a binary result (female or male).

## 3. Deep Learning Neural Networks

### 3.1. Convolutional Neural Network

Convolutional neural networks (CNNs) draw inspiration from the visual cortex, where small cell regions are sensitive to specific visual field areas. These networks blend elements of biology, mathematics, and computer science, becoming pivotal innovations in computer vision and artificial intelligence (AI). The turning point for CNNs came in 2012 when Alex Krizhevsky et al. [33] utilized an 8-layer CNN (5 convolutional, 3 fully-connected) to win the ImageNet competition [34], significantly reducing the classification error rate from 25.8% (in 2011) to 16.4% (in 2012), marking a remarkable advancement at the time. This model (of 63 million parameters), referred to as *AlexNet*, has since become influential in the field.

Simonyan and Zisserman [35] from the University of Oxford crafted a 19-layer (16 conv., 3 fully connected) CNN named the VGG-19 model [35,36]. This model strictly employs 3 × 3 filters with stride and pad of 1, alongside 2 × 2 max-pooling layers with stride 2. VGG Net achieved a remarkable error rate of 7.3% by utilizing small filters in all convolutional layers, underscoring the importance of deep layer networks for effective hierarchical representation of visual data. The VGG-19 model (143 million parameters) ranked 2nd in classification and 1st in localization in ILSVRC 2014.

The ResNet-50 model [37] encompasses 101 layers, featuring 33 three-layer *residual* blocks in addition to input and output layers. These identity connections facilitate learning incremental or residual representations, enabling smooth back-propagation. ResNet-101 employs 3 × 3 filters with a stride of 2 and 3 × 3 max-pooling layers with a stride of 2. Notably, ResNet-101 secured 1st place in ImageNet classification at ILSVRC 2015, addressing the training challenges encountered in deep networks (more than 30 layers), such as vanishing gradients.

Recently, an Xception [38] (Extreme Inception) network architecture has been proposed on a hypothesis: the mapping of cross-channel correlations and spatial correlations in the feature maps of CNNs can be entirely decoupled. Thus, the Inception modules can be replaced with depthwise separable convolutions. The Xception architecture has 36 convolutional layers forming the feature extraction base of the network. For image classification, the convolutional base will be followed by a logistic regression layer. Optionally, one may insert fully connected layers before the logistic regression layer. The 36 convolutional layers are structured into 14 modules, all of which have linear residual connections around them, except for the first and last modules. Compared to Inception V3, Xception has a similar parameter count and shows small gains in classification performance on the ImageNet dataset.

As used in the Xception model, a depthwise separable convolution, commonly called *separable convolution* in deep learning frameworks such as TensorFlow/Keras, consists of a depthwise convolution, i.e., a spatial convolution performed independently over each channel of an input, followed by a *pointwise convolution*, i.e., a 1 × 1 convolution, projecting the channel's output by the depthwise convolution onto a new channel space. The scenario of separable convolution plus pointwise convolution can significantly reduce the load of

convolutional computation in contrast with a regular 2D or 3D convolution layer, thus speeding up the CNN model training and inferring process.

### 3.2. Video Vision Transformer

Transformers were primarily developed and utilized for tasks in natural language processing (NLP), exemplified by language models like BERT (Bidirectional Encoder Representations from Transformers) [39]. Transformers establish connections between pairs of input tokens, such as words in NLP, through a mechanism known as *attention*. However, this approach becomes increasingly computationally intensive as the number of tokens grows. In dealing with images, the fundamental unit of analysis shifts to the pixel. Yet, computing relationships between every pair of pixels becomes prohibitively resource-intensive in terms of memory and computation.

To tackle this challenge, Vision Transformers (ViTs) calculate relationships among smaller image regions, typically $16 \times 16$ pixels, thus reducing computational demands. These regions, along with positional embeddings, form a sequence. The embeddings represent learnable vectors. Each region is vectorized and then multiplied by an embedding matrix. The resulting sequence, along with positional embeddings, is then inputted into the transformer for further processing. The Video ViT (ViViT) model adds an extra step called a video tube (cube), which organizes frames by height by width (e.g., $6 \times 9 \times 8$), along with positional embedding. The remainder of the ViViT process remains the same as ViT.

Self-attention is a common technique applied in the vision transformer model. It involves creating three vectors from each of the encoder's input vectors (in the NLP case, the embedding of each word). For each word, a Query vector, a Key vector, and a Value vector are created by multiplying the embedding by three matrices trained during the training process. Multi-headed attention enhances the model's ability to focus on different positions by providing the attention layer with multiple representation subspaces. This is achieved by utilizing multiple sets of Query/Key/Value weight matrices, where each set represents a different representation subspace. For example, a transformer might employ eight attention heads, each consisting of eight sets of weight matrices for each encoder/decoder. These sets are initialized randomly and then used post-training to project input embeddings or vectors from lower encoders/decoders into distinct representation subspaces.

### 3.3. Recurrent Neural Network

Recurrent neural networks (RNNs) are crafted to harness sequential data. Unlike traditional neural networks, which assume independence between inputs and outputs, RNNs acknowledge the significance of dependencies, which is particularly evident in tasks like NLP. For instance, predicting the subsequent word in a sentence benefits from contextual knowledge about preceding words. RNNs earn their *recurrent* designation because they execute the same operation for each sequence element, with output reliant on prior computations. Conceptually, RNNs possess a "memory" that retains information from previous calculations. While theoretically capable of leveraging information from arbitrarily long sequences, practical implementations often limit consideration to a few preceding steps.

RNNs have demonstrated significant success across various NLP tasks and applications involving temporal signals [40]. Among RNN variants, long short-term memory (LSTM) networks stand out, excelling in capturing long-term dependencies. LSTMs, akin to RNNs, employ a distinct approach to compute the hidden state. In LSTMs, memories, termed cells, act as "black boxes" that take the previous (hidden) state, $s_{t-1}$, and current input, $x_t$, as inputs. These cells autonomously decide what information to retain or discard

from memory before combining the previous state, current memory, and input. Notably, LSTM units are highly effective at capturing long-term dependencies.

In certain applications, such as weather prediction, modeling temporal evolution (e.g., HR changes over time) via recurrence relations (e.g., LSTM) is desirable. Similarly, in vital sign estimation, capturing physiological signal variation over time (e.g., HR fluctuations) is essential. Simultaneously, efficient extraction of spatial features (e.g., signal variation with location) is crucial, often accomplished using convolutional filters. Therefore, an ideal architecture integrates both recurrent and convolutional mechanisms, resulting in convolutional LSTM (ConvLSTM) layers.

### *3.4. Hybrid Models*

Given facial video clips as input data (30 × 288 × 256 × 3) and six continuous output variables (HR, SpO2, SBP, DBP, sex, age), a typical neural network (NN) model comprises several key blocks: Input, Stem, Body, Head, and Output. As outlined in Table 2, the Stem block comprises a single convolutional layer, reducing the input frame size by ¼. The "time_distr" operator applies the same convolutions to all 30 frames. The Head block includes two fully connected layers, with separate heads for each output. All six outputs are then concatenated into a vector output. Table 2 also presents two Body blocks, ViViT (Mdl_a) and ConvLSTM (Mdl_b), which can be concatenated into a hybrid model (HMdl_ab) or utilized separately as two models (Mdl_a, Mdl_b) by excluding the "Concat" operator.

**Table 2.** Hybrid Model (HMdl_ab) architecture combining 2 NN models: ViViT and ConvLSTM. Normalization and Dropout layers are omitted. The batch size is omitted in the "Output Shape" column. time_distr = time_distributed.

| **HMdl_ab: ViViT ⊕ ConvLSTM, 1072.0 M Paras** | | | |
|---|---|---|---|
| **Layer (Type)** | **Output Shape** | **Layer (Type)** | **Output Shape** |
| Frame_Input | (30, 288, 256, 3) | | |
| **Stem**: time_distr (Conv2D): | | | |
| (filters, kernel_size, strides) = | | | |
| (64, (4, 4), (4, 4)) | (30, 72, 64, 64) | | |
| **ViViT** (Mdl_a): Tubelet_Embedding: | | **ConvLSTM** (Mdl_b): | |
| (Conv3D → 30 × 12 × 16 patches) | | (filters, kernel_size, strides) = | |
| (Conv3D → 30 × 3 × 4 patches) | (360, 768) | (128, (3, 2), (3, 2)) | (30, 24, 32, 128) |
| Add (Positional_Encoder) | (360, 768) | (256, (2, 2), (2, 2)) | (30, 12, 16, 256) |
| 8 × Attn_FF: | | (512, (2, 2), (2, 2)) | (30, 6, 8, 512) |
| MultiHeadAttention: | | (768, (2, 2), (2, 2)) | (30, 3, 4, 768) |
| (heads = 8, key_dim = 64) | | | |
| Feed_Forward_Net: | | | |
| (768 → 3072 → 768) | | | |
| Add (Attn_FF_Norm) | (360, 768) | | |
| Flatten | (276480) | Flatten | (276480) |
| Concat | (552960) | | |
| 6 × **Head**: | | | |
| Dense | (256) | | |
| Dense | (1) | | |
| Concat (Output) | (6) | | |

The ViViT (Mdl_a) Body block comprises Tublet embedding (two 3D convolutional layers for patch embedding), positional encoder, multi-head attention, and feed-forward network (FFN). We opted for 1 frame in Tublet embedding to maintain high temporal resolution, 8-head attention, 64 key dimensions, 768 embedding vector dimensions, and 8 FFN

layers. A total of 360 embedding vectors (768 dimensions each) result from 30 temporal frames and 12 spatial patches.

The ConvLSTM (Mdl_b) Body block comprises 4 ConvLSTM layers with specified numbers of filters, convolutional kernel size, and strides. Subsequently, 12 spatial patches over 30 frames are flattened to preserve both spatial and temporal resolution in the feature space.

Table 3 showcases a hybrid model (HMdl_abc) consisting of three Body blocks: ViViT (Mdl_a), ConvLSTM (Mdl_b), and ResCNN (Mdl_c). The ResCNN (Mdl_c) Body block comprises five time-distributed (per frame) convolutional layers, each featuring a residual block. The numbers of filters, convolutional kernel size, and strides are specified for each convolutional layer, followed by a "flatten" function at the end.

**Table 3.** Hybrid Model (HMdl_abc) architecture combining 3 NN models: CNN_Res, ViViT, and ConvLSTM. Normalization and Dropout layers are omitted. The batch size is omitted in the "Output Shape" column. time_distr = time_distributed.

| **HMdl_abc: ViViT ⊕ ConvLSTM ⊕ ResCNN, 1569.2 M Paras** | | | |
|---|---|---|---|
| **Layer (Type)** | **Output Shape** | **Layer (Type)** | **Output Shape** |
| Frame_Input | (30, 288, 256, 3) | | |
| **Stem**: time_distr (Conv2D): (filters, kernel_size, strides) = (64, (4, 4), (4, 4)) | (30, 72, 64, 64) | | |
| **ViViT** (Mdl_a): (from Hybrid_Model_1) | **ConvLSTM** (Mdl_b): (from Hybrid_Model_1) | **ResCNN** (Mdl_c): time_distr (Conv2D): (filters, kernel_size, strides) = | |
| | | (64, (7, 7), (1, 1)) | (30, 72, 64, 64) |
| | | (128, (5, 5), (3, 2)) | (30, 24, 32, 128) |
| … | … | (256, (3, 3), (2, 2)) | (30, 12, 16, 256) |
| | | (512, (3, 3), (2, 2)) | (30, 6, 8, 512) |
| | | (768, (3, 3), (2, 2)) | (30, 3, 4, 768) |
| Flatten (276480) | Flatten (276480) | Flatten | (276480) |
| Concat (ViViT, convLSTM, ResCNN) | (829440) | | |
| 6 × **Head**: | | | |
| Dense | (256) | | |
| Dense | (1) | | |
| Concat (Output) | (6) | | |

### *3.5. Nine Neural Network (NN) Models*

Utilizing three base models, ViViT (Mdl_a), ConvLSTM (Mdl_b), and ResCNN (Mdl_c), we can combine them to form four hybrid models: Mdl_ab, Mdl_ac, Mdl_bc, and Mdl_abc (refer to Tables 2 and 3). In addition, two commonly used CNN models, ResNet-50 and Xception, are time-distributed to each frame in conjunction with the (6×) Head layer (as outputs). The inputs of these two CNN models are the facial video clips ($30 \times 288 \times 256 \times 3$, no Stem layer applied). Consequently, a total of nine NN models are considered for comparison in our experiments. The parameter sizes of these models are provided in Tables 2–4.

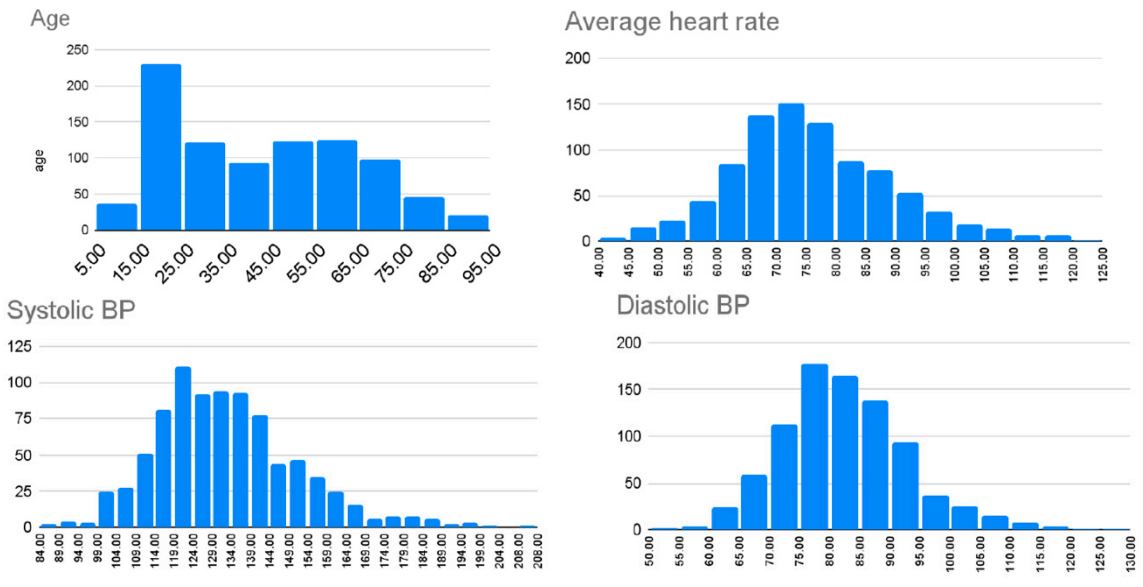
**Table 4.** The root mean square errors (RMSEs) and the root mean square deviations (RMSDs) of vital signs (HR, SpO2, SBP, DBP) prediction varying with nine deep learning NN models: Trained on the data samples from 90% of subjects and Tested on the samples from 10% subjects from the VVD-Large dataset [27]. In the left column, the model parameters and its output features (to Heads) are presented.

| Model\Vital Sign | HR | SpO2 | SBP | DBP | Sex | Age |
|---|---|---|---|---|---|---|
| ResNet-50, 118.1 M 61,440 | 11.11 (17.16) | 1.54 (1.15) | 9.62 (22.14) | 6.51 (13.95) | 0.06 (0.49) | 6.27 (20.95) |
| Xception, 115.4 M 61,440 | 7.90 (15.32) | 1.24 (1.89) | 6.11 (20.74) | 3.86 (12.35) | 0.04 (0.49) | 3.60 (20.81) |
| Mdl_a, 552.0 M ViViT: 276,480 | 7.02 (12.14) | 0.99 (1.52) | 6.69 (17.16) | 3.72 (10.10) | 0.02 (0.50) | 4.38 (19.18) |
| Mdl_b, 520.0 M ConvLSTM: 276,480 | 7.12 (10.51) | 1.04 (1.19) | 5.92 (15.20) | 3.75 (8.61) | 0.02 (0.50) | 4.40 (17.94) |
| Mdl_c, 497.2 M ResCNN: 276,480 | 6.60 (12.61) | 0.90 (1.54) | 5.61 (17.68) | 3.34 (10.72) | 0.01 (0.50) | 3.65 (19.74) |
| HMdl_ab, 1072.0 M 552,960 | 6.49 (11.50) | 0.93 (1.35) | 5.18 (16.30) | 3.05 (9.50) | 0.02 (0.50) | 3.01 (18.49) |
| HMdl_ac, 1049.1 M 552,960 | 6.33 (12.64) | 0.88 (1.57) | 4.71 (17.91) | 2.80 (10.67) | 0.03 (0.50) | 2.72 (19.68) |
| HMdl_bc, 1017.2 M 552,960 | 6.56 (11.37) | 0.97 (1.38) | 4.94 (16.34) | 2.89 (9.55) | 0.02 (0.50) | 3.23 (18.56) |
| HMdl_abc, 1569.2 M 829,440 | 6.42 (11.67) | 0.88 (1.43) | 4.84 (16.56) | 2.80 (9.67) | 0.00 (0.50) | 2.56 (18.96) |

## 4. Experimental Results

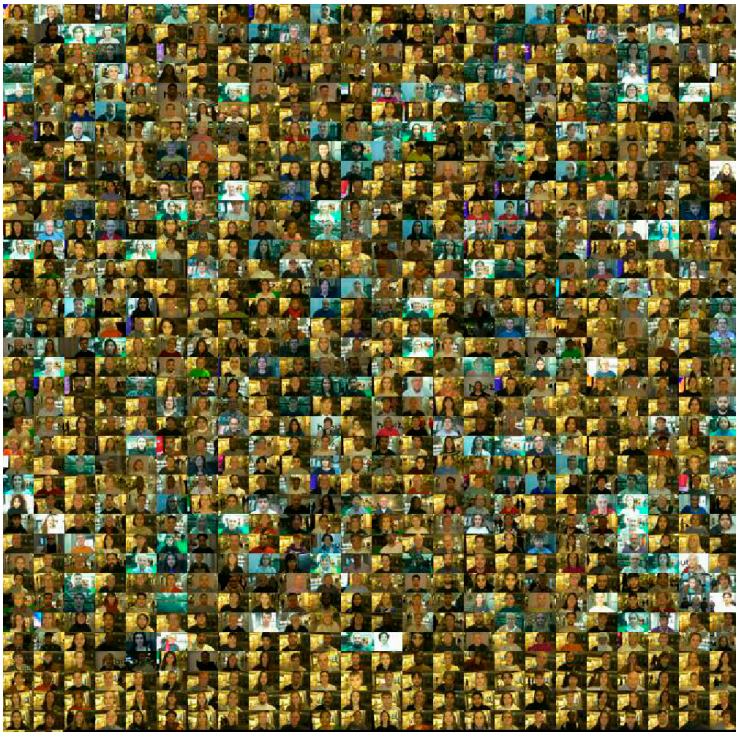
### *4.1. Vital Video Dataset Description*

We assembled a sizable dataset known as the Vital Video Dataset (VVD) [27], comprising 891 participants, illustrated in Figure 1. This dataset encompasses 459 female and 432 male individuals. For each participant, we captured two uncompressed 30 s videos, synchronized PPG waveforms, and a single blood pressure measurement. Additionally, gender, age, and skin color were recorded for every participant.




(**a**)

**Figure 1.** *Cont.*



(**b**)

**Figure 1.** Illustration of the VVD-Large dataset [27] (891 subjects): (**a**) the distributions of the variables, (**b**) mosaic of the facial videos (851 = 37 × 23 faces shown).

During the recording process, participants were instructed to maintain a forward gaze and keep their fingers as steady as possible, preferably resting on the knee. Once the PPG waveform stabilized, a 30-second video was recorded simultaneously with the PPG waveform. Furthermore, heart rate and blood oxygen saturation level values calculated by the pulse oximeter were recorded.

Following the initial recording, another 30-second session commenced, during which BP measurements were obtained. PPG waveform, HR, and SpO2 values were also captured during this session. It is noteworthy that measuring BP during the recording, rather than before or after, is crucial as the measurement process itself can influence BP values.

The details of the VVD dataset are described as follows:

- A total of 891 participants (balanced gender/age distribution including all skin tones);
- Uncompressed 1920 × 1200, 30 FPS facial videos (see Figure 1b);
- Synchronized PPG ground truth;
- Blood pressure spot measurement collected during video recording;
- Sex and age recorded.

In our experiment, we employed a compressed version of the VVD dataset, named the VVD-Large dataset, comprising 891 participants. This dataset encompasses a total of 90 GB of compressed videos at a bitrate of 13 Mbps. The distributions of age, averaged HR, and BP values are depicted in Figure 1a. The attributes available in the VVD-Large dataset [27] are listed below, where the first reading of each attribute in the first example is provided as an example.

i. Heart rate (HR): one reading per second in beats per minute, e.g., 72;
ii. Blood oxygen saturation level (SpO2): one reading per second in percentage (≤100%), e.g., 98;
iii. Blood pressure (BP): one reading per second, including systolic pressure and diastolic pressure, in millimeters of mercury (mm Hg), e.g., 126/66;

iv. Subject sex: e.g., M;
v. Subject age: e.g., 22;
vi. PPG waveform: 56 readings per second in percentage (not used in this study), e.g., 35;
vii. Subject skin color: labeled as 1 (White)–6 (dark brown/black) according to Fitzpatrick skin types (not used in this study), e.g., 4.

Each participant is associated with two video files, each containing 30 s facial videos. As one heartbeat cycle typically lasts less than one second (i.e., HR typically exceeds 60 bpm or 1 bps), we clip 30 frames (i.e., 1 s of 30 FPS video) as the sequence of input images for our NN models. From each of the two video files for a participant, nine 30-frame clips are randomly extracted. It is ensured that there are no overlapping (repeatedly used) frames in the nine clips from the same participant. During data splits for training and testing, we ensure that multiple clips from one participant are allocated to the same subset, either the training or testing subset.

All faces are detected and extracted from each frame, resized to 288 × 256 pixels, and then standardized and normalized (intensity stretched) as per Section 2. The normalized facial video clips serve as the input data, while the outcome variables (outputs) consist of four vital signs, HR, SpO2, Systolic BP, and Diastolic BP, along with sex and age. Ground-truth values corresponding to these six outcome variables are normalized to [0, 1] for training, and the NN model's prediction values are scaled back to their original range. A thresholding process is applied for sex prediction.

Given the model's complexity and the large video data, training an NN model is time consuming (e.g., 1018 s/epoch for HMdl_abc). Instead of 10-fold cross-validation, 90% of the VVD-Large data (from 802 subjects) is utilized for training, while 10% (from 89 subjects) is reserved for testing. All reported results in Table 4 are obtained from the testing dataset. An illustration of one hybrid model training is presented in Figure 2, demonstrating the convergence of the train loss and validation loss after 30 epochs. All trainings are halted after 64 epochs.

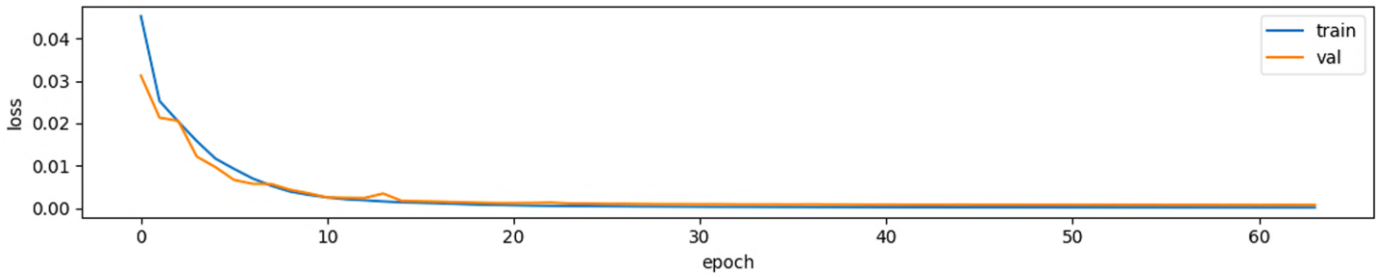



**Figure 2.** Train and validation (10% of train dataset) loss over 64 epochs (HMdl_abc, 709.7 M, 16 h of training).

### *4.2. Nine NN Model Performance*

The performance of the NN model prediction is evaluated using root mean square error (RMSE) and root mean square deviation (RMSD) by comparing the prediction values with the ground-truth values. A smaller RMSE indicates better prediction accuracy. It is worth noting that RMSD is influenced by the range (distribution) of each vital sign variable and the prediction performance. When comparing RMSD values between different NN models (but not across different vital signs), smaller values indicate better performance.

Table 4 showcases the performance of nine NN models in predicting four vital signs (HR, SpO2, SBP, DBP), as well as sex and age. The ViViT (Mdl_a) model incorporates both temporal and spatial analysis into a single model, applying attention mechanisms to both time patches (among frames) and spatial patches (within a frame). All four hybrid models (HMdl_ab, _ac, _bc, and _abc) exhibit smaller RMSEs compared to the three base models

(Mdl_a, _b, and _c). Overall, the hybrid model, HMdl_ac and HMdl_abc, demonstrates the best performance, suggesting that combining ViViT with ResCNN and ConvLSTM further enhances spatial analysis capabilities atop the already high temporal (between frames) analysis power in the Mdl_a model.

As shown in Table 4, the performance of the Xception model is better than the ResNet-50 model. However, both are poorly performed compared to the rest of the seven customized models. Therefore, the rest of the discussion, including score fusion, is concentrated on the seven proposed models.

To visualize the prediction performance, histograms of prediction values and ground-truth values are depicted in Figure 3. These plots illustrate the distributions of four vital signs and ages. In the bottom-left subplot of Figure 3, if a prediction bar (orange) is lower than its neighboring ground-truth bar (green), it indicates that the average prediction is smaller than the average ground-truths within that bin (a range of vital sign).

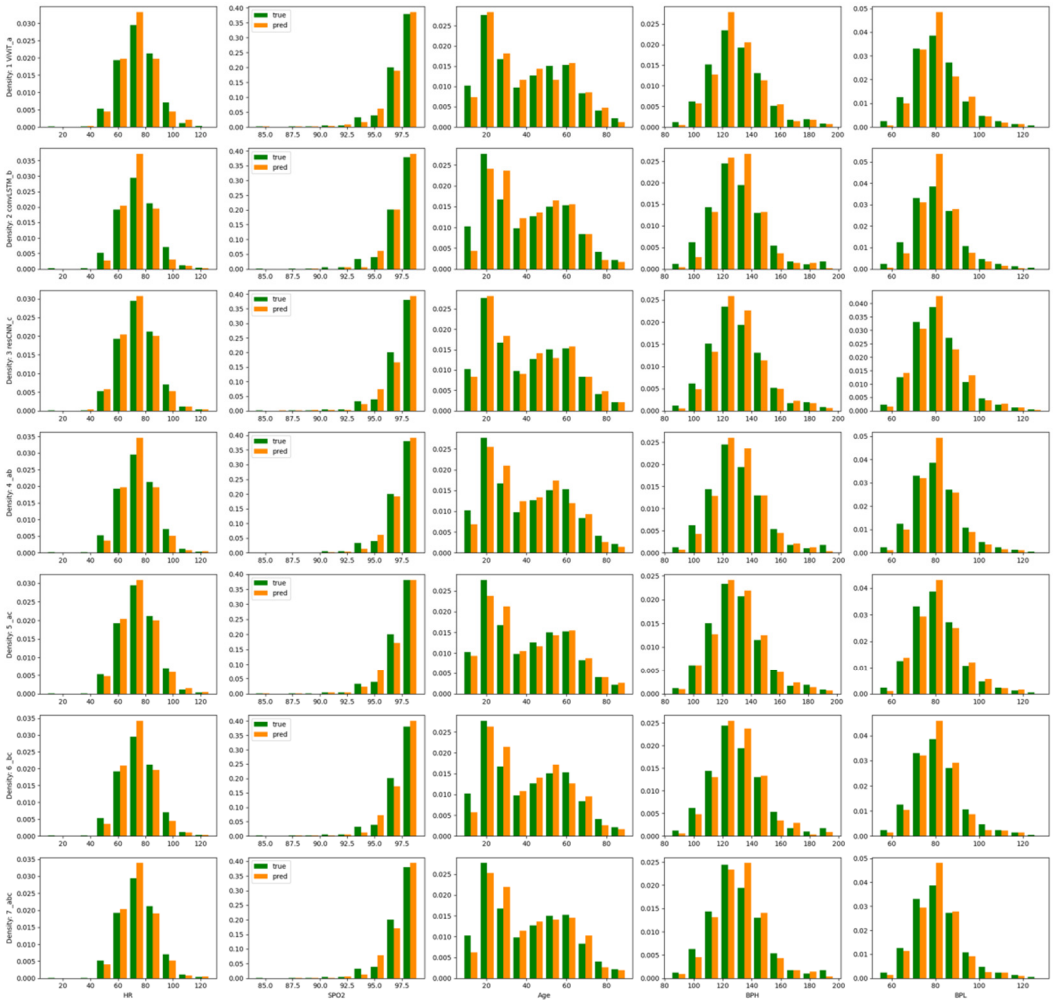



**Figure 3.** Vital sign predictions (5 columns: HR, SpO2, Age, BPH/DBP, BPL/SBP) comparisons of seven models (7 rows: Mdl_a, _b, _c, HMdl_ab, _ac, _bc, _abc): true (ground truth in green color) vs. prediction (in orange color).

Among these outcome variables, sex and age are considered static variables, meaning they do not vary over time within the recording duration. SpO2 is a slowly varying variable. However, HR and BPs are dynamic variables, exhibiting rapid or substantial fluctuations within 30–60 s. To accurately predict fast-varying variables, temporal data analysis with high resolution is crucial. The ViViT model (Mdl_a) offers temporal analysis (via attention

mechanism) with a resolution of 30 frames. While two hybrid models (HMdl_ac and HMdl_abc) demonstrate smaller errors in predicting SpO2, DBP, Sex, and Age, they exhibit relatively larger errors in estimating HR and SBP. Achieving higher temporal resolution may be feasible by employing a higher frame rate camera (e.g., 90 FPS, 300 FPS) and implementing a novel temporal analysis model. Hopefully, these measures will help reduce the errors in predicting highly dynamic vital signs like HR and SBP.

### 4.3. Score Fusion Performance

To enhance the performance of vital sign estimation, score fusion methods can be employed, integrating multiple scores from different NN models, as delineated in Table 4. Various types of score-fusion methodologies exist, including arithmetic fusion (e.g., average, majority vote) [41], regression-based fusion, and density-based fusion (e.g., Gaussian mixture model) [42,43]. For this study, three regression-based score fusion methods were chosen: Mean fusion (baseline), Support Vector Machine (SVM) [44,45], Random Forest (RF), and Gradient Boosting. The multiple scores are amalgamated into feature vectors and subsequently inputted into a classifier for training (utilizing labeled score vectors) or testing.

In our study, the SVM method utilized a Gaussian kernel function and a one-versus-one coding design, employing seven binary learners for the corresponding seven scores. The RF model involved training an ensemble of 100 classification trees, each with a depth of 40, using the complete training dataset. At each decision split, a random subset of predictors (scores) was employed to maximize the gain of the split criterion across all possible splits of the predictors. The final regression was derived by aggregating the results from all the trees in the ensemble.

Ten-fold cross-validation was implemented in score fusion experiments, and the root mean square errors (RMSEs) of score fusion were calculated by combining all ten-fold testing scores. The scores for fusion, a feature vector of seven dimensions, consist of the scores from seven NN models (see Table 5).

**Table 5.** The RMSEs of score fusion using 7 scores from 7 NN models.

| Fusion\Vital Sign | HR | SpO2 | SBP | DBP | Sex | Age |
|---|---|---|---|---|---|---|
| Mean Fusion | 6.18 | 0.87 | 4.65 | 2.70 | 0.01 | 2.60 |
| SVM Regr. | 5.86 | 0.98 | 7.66 | 3.86 | 0.09 | 3.52 |
| Gradient Boosting | 5.53 | 0.89 | 4.38 | 2.34 | 0.00 | 2.38 |
| RF Regr. (100, 40) | 5.55 | 0.87 | 4.07 | 2.21 | 0.00 | 2.30 |

The mean fusion simply averages all scores. Except for the SVM regression, the rest of the score fusion methods (including mean fusion) performed better than any single model. The RF regression method is the best model in terms of the smallest RMSEs.

### 4.4. Time Costs of Nine NN Models

All models were implemented using Python 3.9 and TensorFlow 2.11 and executed in JupyterLab (Version 3.6.1) on a DELL Precision workstation with the following specifications: Intel Xeon 5220R 24-Core CPU clocked at 2.2 GHz, 256 GB of RAM, 8 TB hard disk, running Ubuntu 20.04. The workstation was equipped with an NVIDIA Dual RTX A6000 Graphics Board (with NV Link) featuring 48 GB of video memory and 10,752 CUDA cores for each board. However, all training and testing procedures were conducted using only one A6000 board. The time costs of training and testing for each of the nine NN models are detailed in Table 6. There is no doubt that the ResNet-50 and Xception models are fast in training and testing due to the small parameter size. Notably, the hybrid model (HMdl_abc)

required the longest time for both training and testing. Nevertheless, the inference time of 56.8 milliseconds per sample (consisting of 30 frames) is relatively rapid, enabling the machine to process approximately 18 such requests in a single second.

**Table 6.** Model parameters, dimension of the output features (to Heads), training time (including validation time, seconds per epoch), and testing time (milliseconds per sample) varying with nine NN models tested on the VVD-Large dataset [27]. Time costs (testing on one Nvidia A6000 48 GB-GPU board) slightly change at different runs due to data caching and optimization.

| Metric\NN Model | ResNet-50 | Xception | Mdl_a | Mdl_b | Mdl_c | HMdl_ab | HMdl_ac | HMdl_bc | HMdl_abc |
|---|---|---|---|---|---|---|---|---|---|
| # Parameters | 118.1 M | 115.4 M | 552.0 M | 520.0 M | 497.2 M | 1072.0 M | 1049.1 M | 1017.2 M | 1569.2 M |
| # Features | 61,440 | 61,440 | 276,480 | 276,480 | 276,480 | 552,960 | 552,960 | 552,960 | 829,440 |
| Training time (s/epoch) | 235 | 256 | 348 | 420 | 396 | 519 | 477 | 566 | 1018 |
| Testing time (ms/sample) | 8.3 | 9.3 | 23.5 | 23.5 | 23.5 | 23.5 | 23.5 | 23.5 | 43.1 |

## 5. Discussion and Conclusions

This study introduces a deep learning framework aimed at directly predicting vital signs from facial videos. Comparing nine NN models, the hybrid model, which integrates CNN, convLSTM, and ViViT blocks, emerges as the top performer. The small RMSEs associated with the hybrid model's predictions indicate its superior capability in extracting temporospatial features from facial videos. Through experiments, we observed the hybrid model's efficacy in predicting vital signs such as HR, SpO2, SBP, and DBP, along with demographic variables like sex and age.

Facial images can be easily captured using common devices like smartphones, laptops, or web cameras, eliminating the need for specialized equipment like vital sign monitors. This implementation offers an economical and convenient solution for vital sign monitoring, particularly beneficial for elderly individuals or during outbreaks of contagious diseases like COVID-19.

Due to the large model and big dataset but limited GPU resources, we did not conduct 10-fold cross-validation experiments. Instead, we randomly chose one-fold to do the test. Thus, error analysis cannot be performed in this study. The VVD-Large dataset does not include any ground truths for body temperature and respiration rate, although the proposed hybrid model has the capacity to integrate body temperature and respiration rate prediction.

The proposed models were only validated on one dataset, although this dataset is fairly large and has already considered the balances of age, gender, and race. However, the dataset (size, representation, diversity, balance, processing) plays an important role in any data-driven model, including ours hereby. Thus, we will continuously validate and improve our models on larger datasets.

The data preprocessing was completed prior to any model training and testing in our experiments. For real-time applications, facial detection and data normalization can be implemented either on the client side (e.g., in a smartphone app) or on the server side (where the hybrid model performs predictions). A trained hybrid model can be further improved by applying reinforcement learning methods.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data may be available upon request from the authors in [27]. The source code and pre-trained models proposed in this paper are available for download at https://github.com/yfzheng2000/Vital_Sign_Pred (accessed on 12 January 2025).

**Acknowledgments:** The author would like to thank the dataset provider who shared his precious VVD-Large dataset [27].

**Conflicts of Interest:** The author declares no conflicts of interest.

## References

1. Bousefsaf, F.; Maaoui, C.; Pruski, A. Peripheral vasomotor activity assessment using a continuous wavelet analysis on webcam photoplethysmographic signals. *Bio. Med. Mater. Eng.* **2016**, *27*, 527–538. [CrossRef] [PubMed]
2. Jeong, I.C.; Finkelstein, J. Introducing contactless blood pressure assessment using a high speed video camera. *J. Med. Syst.* **2016**, *40*, 77. [CrossRef] [PubMed]
3. Shao, D.; Liu, C.; Tsow, F.; Yang, Y.; Du, Z.; Iriya, R.; Yu, H.; Tao, N. Noncontact monitoring of blood oxygen saturation using camera and dual-wavelength imaging system. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1091–1098. [CrossRef] [PubMed]
4. Poh, M.Z.; McDuff, D.J.; Picard, R.W. Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Trans. Biomed. Eng.* **2011**, *58*, 7–11. [CrossRef] [PubMed]
5. Poh, M.Z.; McDuff, D.J.; Picard, R.W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express* **2010**, *18*, 10762–10774. [CrossRef]
6. Chen, W.; McDuff, D. Deepphys: Video-based physiological measurement using convolutional attention networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 349–365.
7. Hu, M.; Qian, F.; Wang, X.; He, L.; Guo, D.; Ren, F. Robust heart rate estimation with spatial–temporal attention network from facial videos. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 639–647. [CrossRef]
8. Lokendra, B.; Puneet, G. And-rppg: A novel denoising-rppg network for improving remote heart rate estimation. *Comput. Biol. Med.* **2022**, *141*, 105146. [CrossRef] [PubMed]
9. Yin, R.N.; Jia, R.S.; Cui, Z.; Sun, H.M. Pulsenet: A multitask learning network for remote heart rate estimation. *Knowl.-Based Syst.* **2022**, *239*, 108048. [CrossRef]
10. Li, B.; Zhang, P.; Peng, J.; Fu, H. Non-contact ppg signal and heart rate estimation with multi-hierarchical convolutional network. *Pattern Recogn.* **2023**, *139*, 109421. [CrossRef]
11. Luo, H.; Yang, D.; Barszczyk, A.; Vempala, N.; Wei, J.; Wu, S.J.; Zheng, P.P.; Fu, G.; Lee, K.; Feng, Z.P. Smartphone-based blood pressure measurement using transdermal optical imaging technology. *Circ. Cardiovasc. Imaging* **2019**, *12*, e008857. [CrossRef] [PubMed]
12. Wu, B.F.; Chiu, L.W.; Wu, Y.C.; Lai, C.C.; Chu, P.H. Contactless blood pressure measurement via remote photoplethysmography with synthetic data generation using generative adversarial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2130–2138.
13. Uchi, K.; Miyazaki, R.; Cardoso, G.C.; Ogawa-Ochiai, K.; Tsumura, N. Remote estimation of continuous blood pressure by a convolutional neural network trained on spatial patterns of facial pulse waves. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2139–2145.
14. Song, R.; Chen, H.; Cheng, J.; Li, C.; Liu, Y.; Chen, X. PulseGAN: Learning to Generate Realistic Pulse Waveforms in Remote Photoplethysmography. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 1373–1384. [CrossRef]
15. de Haan, G.; Jeanne, V. Robust Pulse Rate From ChrominanceBased rPPG. *IEEE Trans. Biomed. Eng.* **2013**, *60*, 2878–2886. [CrossRef]
16. Yu, Z.; Peng, W.; Li, X.; Hong, X.; Zhao, G. Remote Heart Rate Measurement From Highly Compressed Facial Videos: An End-to-End Deep Learning Solution With Video Enhancement. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.
17. Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Torr, P.H.S.; Zhao, G. PhysFormer: Facial Video-Based Physiological Measurement with Temporal Difference Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
18. Hu, M.; Qian, F.; Guo, D.; Wang, X.; He, L.; Ren, F. ETArPPGNet: Effective Time-Domain Attention Network for Remote Heart Rate Measurement. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 2506212. [CrossRef]
19. Gideon, J.; Stent, S. The Way to my Heart is through Contrastive Learning: Remote Photoplethysmography from Unlabelled Video. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
20. Hsu, G.-S.J.; Xie, R.-C.; Ambikapathi, A.; Chou, K.-J. A deep learning framework for heart rate estimation from facial videos. *Neurocomputing* **2020**, *417*, 155–166. [CrossRef]

21. Omer, O.A.; Salah, M.; Hassan, L.; Abdelreheem, A.; Hassan, A.M. Video-based beat-by-beat blood pressure monitoring via transfer deep-learning. *Appl. Intell.* **2024**, *54*, 4564–4584. [CrossRef]
22. Cheng, C.-H.; Yuen, Z.; Chen, S.; Wong, K.-L.; Chin, J.-W.; Chan, T.-T.; So, R.H.Y. Contactless Blood Oxygen Saturation Estimation from Facial Videos Using Deep Learning. *Bioengineering* **2024**, *11*, 251. [CrossRef]
23. Jaiswal, K.B.; Meenpal, T. Heart rate estimation network from facial videos using spatiotemporal feature image. *Comput. Biol. Med.* **2022**, *151*, 106307. [CrossRef] [PubMed]
24. Lin, B.; Tao, J.; Xu, J.; He, L.; Liu, N.; Zhang, X. Estimation of vital signs from facial videos via video magnification and deep learning. *iScience* **2023**, *26*, 107845. [CrossRef] [PubMed] [PubMed Central]
25. Jinsoo, P.; Kwangseok, H. Facial Video-Based Robust Measurement of Respiratory Rates in Various Environmental Conditions. *J. Sens.* **2023**, *2023*, 9207750. [CrossRef]
26. Zheng, Y.; Wang, H.; Hao, Y. Mobile application for monitoring body temperature from facial images using convolutional neural network and support vector machine. In *Mobile Multimedia/Image Processing, Security, and Applications 2020*; Proceedings SPIE 11399; SPIE: Bellingham, WA, USA, 2020; p. 113990B. [CrossRef]
27. Toye, P.J. Vital Videos: A dataset of face videos with PPG and blood pressure ground truths. *arXiv* **2023**, arXiv:2306.11891.
28. Yang, M.H.; Kriegman, D.J.; Ahuja, N. Detecting faces in images: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 34–58. [CrossRef]
29. Viola, P.; Jones, M. Robust real-time object detection. *Int. J. Comput. Vis.* **2001**, *57*, 137–154. [CrossRef]
30. Viola, P.; Jones, M. Rapid Object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. 511–518.
31. Papageorgiou, C.; Oren, M.; Poggio, T. A general framework for object detection. In Proceedings of the Sixth International Conference on Computer Vision, Bombay, India, 7 January 1998; pp. 555–562.
32. Freund, Y.; Schapire, R. A decision theoretic generalization of on-line learning and an application to boosting. In Proceedings of the Computational Learning Theory: Eurocolt'95, Barcelona, Spain, 13–15 March 1995; pp. 23–37.
33. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 1, pp. 1097–1105.
34. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
35. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
36. University of Oxford, Visual Geometry Group. Available online: http://www.robots.ox.ac.uk/~vgg/research/very_deep/ (accessed on 1 November 2024).
37. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
38. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
39. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805v2.
40. Wensel, J.; Ullah, H.; Munir, A. ViT-ReT: Vision and Recurrent Transformer Neural Networks for Human Activity Recognition in Videos. 2022. Available online: https://arxiv.org/pdf/2208.07929.pdf (accessed on 16 August 2022).
41. Kuncheva, L.I. A Theoretical Study on Six Classifier Fusion Strategies. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 281–286. [CrossRef]
42. Prabhakar, S.; Jain, A.K. Decision-level Fusion in Fingerprint Verification. *Pattern Recognit.* **2002**, *35*, 861–874. [CrossRef]
43. Ulery, B.; Hicklin, A.R.; Watson, C.; Fellner, W.; Hallinan, P. *Studies of Biometric Fusion*; NIST Interagency Report; US Department of Commerce, National Institute of Standards and Technology: Gaithersburg, MD, USA, 2006.
44. Burges, C.J. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167. [CrossRef]
45. Hastie, T.; Tibshirani, R.; Friedman, J.H. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: Berlin/Heidelberg, Germany, 2009.