Journal of Informatics and Web Engineering
Vol. 5 No. 2 (June 2026) eISSN: 2821-370X

Intelligent Telemedicine Systems for Contactless Heart Rate Estimation using Deep Learning-based rPPG

Do Duc Thinh[1], Nguyen Duc Manh[2], Nguyen Thi Bich Ngoc[3], Vu Van Huan[4*]
[1,2,4]Faculty of International Education, University of Transport and Communications, No. 3 Cau Giay Street, Lang Thuong Ward, Dong Da District, Hanoi 100000, Vietnam.
[3]Faculty of Information Technology, Sao Do University, No. 24 Thai Hoc 2 Street, Sao Do Ward, Chi Linh City, Hai Duong 170000, Vietnam.
[4]Faculty of Information Technology, Hanoi University of Natural Resources and Environment, No. 41A Phu Dien Street, Phu Dien Ward, Bac Tu Liem District, Hanoi 100000, Vietnam.
**corresponding author: (vvhuan@hunre.edu.vn; ORCiD: 0009-0006-6198-480X)*

Abstract - Over the last ten years, telemedicine has undergone significant developments, from simple communication-based healthcare to smart and data-driven solutions. The solutions are now powered by Artificial Intelligence (AI), Machine Learning (ML), and Internet of Medical Things (IoMT) technologies. Physiological monitoring is traditionally done with contact-based sensors. The sensors include Electrocardiogram (ECG) and pulse oximeters. The sensors have various disadvantages, including hardware costs, difficulty in continuous usage, and discomfort for patients. Photoplethysmography (rPPG), a remote method of physiological monitoring, is a breakthrough technology. The method is used to estimate physiological signals, including Heart Rate (HR), from video streams captured by standard RGB cameras. The paper aims to explore spatial-temporal deep learning frameworks for remote rPPG as part of a standard five-layer telemedicine architecture. We explain a general ML pipeline, along with the benefits of decomposing the spatial and temporal features of images and motion to improve signal extraction against environmental noise. Besides, the paper presents some of the deployment issues, which include motion effects, lighting, as well as bias in the algorithms, specifically with respect to melanin absorption in human skin. The research also presents some of the avenues of future research, specifically with respect to model compression, which will help move from a cloud to an edge device, thus helping to improve the privacy of users.

Keywords—Telemedicine, Remote Photoplethysmography, Electrocardiogram, Heart Rate Estimation, Two-Stream Convolutional Network, Edge AI, Digital Health

Received: 24 February 2026; Accepted: 21 April 2026; Published: 16 June 2026

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2026.5.2.24

Published by MMU Press. URL: journals.mmupress.com/jiwe

1. INTRODUCTION

Throughout the world, the health care industry is in the midst of a massive transformation. This is because the population is aging, and more people than ever have chronic health problems, not to mention a shortage of health care professionals. Continuing to offer telehealth as a primary source of medical care has increased greatly because of the COVID-19 pandemic.

Telemedicine systems were originally designed to offer patients occasional remote consultations using video conferencing. However, these systems have now evolved to intelligent systems where the system collects information, automatically diagnoses the patient, and even conducts risk stratification. Recent studies have revealed that AI is expanding to reach the domain of comprehensive healthcare, for example, using AI to automate the diagnosis of tuberculosis from a chest X-ray [1] and using AI to predict mental wellbeing by deliberately modifying thought patterns [2]. This is because the basis on which these systems are built is referred to as IoMT. IoMT stands for the Internet of Medical Things. The conventional meaning of IoMT refers to wearable sensors, which only come into contact with the body in order to generate data, though they are very accurate. In addition, wearable sensors have limitations for large-scale use, including high distribution costs, battery characteristics, skin irritability (for consumers), and low patient compliance, especially for senior patients.

In order to provide users with more widespread access to Heart Rate (HR) measurements without limitations imposed by these devices using traditional methods, the development of Remote Photoplethysmography (rPPG) was made possible with the use of computer vision technology and deep learning. rPPG allows for continuous measurement of the amount of blood being pumped through our bodies by measuring changes in the colour of human skin when blood enters or leaves. rPPG does this from images taken from colour standard RGB cameras; this permits the use of readily available consumer electronics (e.g., laptops, mobile devices) to provide contactless HR monitoring to the average user [3],[4].

Despite the large potential of using rPPG in telemedicine applications, there are still many challenges to be overcome in order to deploy it in real-world telemedicine environments. The cardiac signal is weak and can be easily overwhelmed by visual noise:

1. Motion artifacts: Head movement and facial expressions result in large pixel intensity changes compared to the small changes caused by blood flow.
2. Lighting conditions: Flickering light, shadows, and changes in lighting colour temperature may degrade the signal quality.
3. Melanin is the main contributor to human skin pigmentation. It has strong light-absorbing capabilities.

Consequently, it is much more difficult to obtain accurate remote rPPG from people with darker skin tones. In order to overcome these problems, this paper provides a comprehensive study of a real-world intelligent web-based telemedicine system using a deep learning spatial-temporal framework.

2. LITERATURE REVIEW

2.1 The Physics of Photoplethysmography (PPG)

The principle behind remote rPPG can be described as mainly based on the characteristics of human tissue and blood [5]. When natural light falls on human tissue, part of it reflects back to the surface, while another part penetrates into the epidermis and dermis layers. In human tissue, different chromophore components absorb different parts of light. These components include melanin, lipids, and particularly haemoglobin in human blood vessels. During the cardiac cycle, systolic and diastolic phases create different pulse changes in human arteries in the microvascular system. These pulse changes create changes in the absorption and reflected light to the sensor. By continuously monitoring these pulse-induced colour changes in human tissue, a waveform can be reconstructed to show the cardiac cycle in a reliable manner. This waveform is called PPG signal [6].

2.2. Evolution of rPPG Algorithms

The development process of rPPG methods can be divided into two stages in principle: traditional methods based on traditional signal processing techniques and more recent data-driven methods. In early studies, it was demonstrated that heart pulses can be retrieved by averaging the green channel in a facial Region of Interest (ROI), because the absorption capacity of oxygenated haemoglobin is the highest in the green band [7]. In addition, some improvements in this area have attempted to decompose the signal from noise, which is often caused by changes in illumination and motion, by means of Blind Source Separation (BSS) techniques, e.g., Independent Component Analysis (ICA) [8] and Principal Component Analysis (PCA). To address this problem of motion noise even more effectively, mathematical models of reflection from the skin were proposed, giving birth to chrominance-based solutions such as CHROM [9] and the Plane Orthogonal to Skin (POS) algorithm [5]. Though these solutions are computationally light, they may not guarantee the quality of the signal under heavy head movement and changes in lighting. This is the reason for the recent trend toward deep learning solutions. Current solutions learn robust spatio-temporal features from raw video frames using 3D Convolutional Neural Networks (3D-CNN) [10] and Vision Transformers (ViT) [11].

2.3. Spatial-Temporal Deep Learning Approaches

However, the main challenge in this area has been finding a delicate balance between computational efficiency and predictive accuracy. To overcome this, recent deep learning architectures have moved in the direction of joint spatial and temporal learning [12]. The basic philosophy behind this is quite simple and intuitive as well. Physiological signals, such as blood volume, and environmental noise, such as rigid head motions, have completely different spatial and temporal behaviours. Hence, recent architectures have moved in the direction of completely separating the learning of spatial appearance and temporal motion of the face, which has helped in effectively filtering out the weak rPPG signal and noise caused by head motions. The work presented in this manuscript is a direct extension of the spatial-temporal learning paradigm. Inspired by recent advances in the area of physiological signal extraction [13], we propose a framework that is practical and useful for telemedicine applications. Instead of relying on complex 3D convolutional layers, we propose the use of spatial attention mechanisms and temporal modulation blocks, which greatly enhance the features learned by the model and at the same time retain the efficiency of the model for seamless web and mobile applications.

3. RESEARCH METHODOLOGY

3.1 Model Architecture

Figure 1 illustrates the high-level architecture of an intelligent telemedicine system for rPPG signal extraction. The proposed deep learning framework replaces traditional heuristic feature engineering with an end-to-end learning strategy for rPPG signal extraction. Specifically, the framework adopts a dual-stream architecture that separates spatial appearance information from temporal motion dynamics into two interacting pathways.

Typically, a Spatial Stream processes a static Mean Frame using a series of 2D Convolutional Neural Network (CNN) blocks. Its main purpose is to learn the spatial topography of the face and create an attention mask denoted as M_{spa}. The mask will have higher probability values for areas with high vascularization, i.e., the forehead and cheeks. At the same time, it will have near-zero weights for areas without pulsatile flow, i.e., eyes, hair, and background.

At the same time, a Temporal Stream will process the changing Difference Frames. Nevertheless, with the aim of reducing the computational cost of the 3D convolution, the most recent frameworks have made use of efficient techniques for modelling the temporal dependencies, including the Temporal Shifting Mechanism (TSM) as well as the channel-wise modulation block. The idea behind the shifting mechanism is that it shifts a fraction of the feature channels over the t dimension. This implies that a fraction of the channels will be shifted to $t-1$, while the remaining fraction will be shifted to $t+1$. This allows regular 2D convolutional layers to have a 3D spatiotemporal receptive field at little or no extra parameter cost. Finally, it has to be noted that all models that belong to this paradigm have employed a late fusion technique. This technique combines Attention Masks for spatial domain, i.e., M_{spa}, with features for temporal domain, i.e., F_{temp}, using the Hadamard product, i.e., $F_{fused} = F_{temp} \otimes M_{spa}$. This mathematical technique ensures that only pulse signals are extracted by the network from valid skin pixels, effectively removing any possible noise that could be introduced by background images and movements.

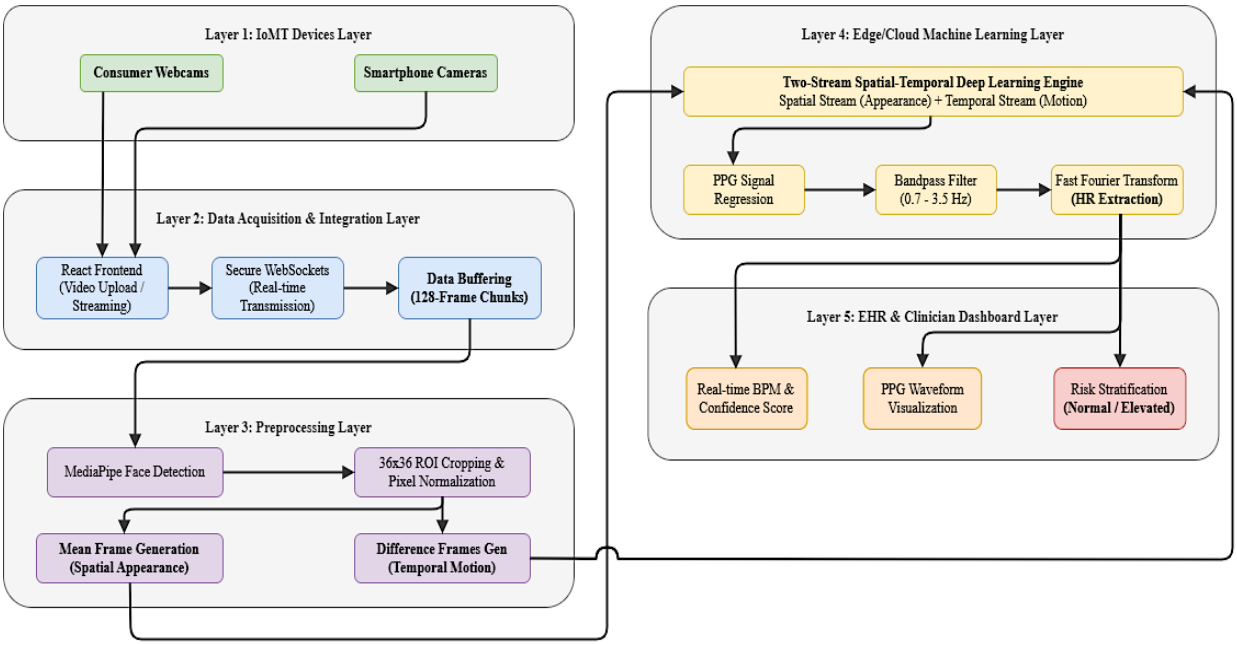

Figure 1. High-Level Architecture of an Intelligent Telemedicine System

In order to ensure that the intelligent rPPG system works effectively in a heterogeneous clinical setting, it must be designed in such a way that it scales well and has a modular design. This way, it can effectively deal with the constant flow of videos and complex deep learning-based tasks without any lag or delay in the results. The framework can easily fit in the traditional five-layer telemedicine architecture.

3.2 IoMT Devices Layer

In contrast, in conventional telemedicine infrastructures, special medical equipment, e.g., pulse oximeters and Holter monitors, have to be transported to the patient. In the proposed system, we have successfully implemented the concept of hardware democratization. In this regard, the main 'sensors' employed in the system are the conventional RGB webcams integrated in the personal computers, tablets, or mobile phones of the patients. By employing conventional consumer electronics as the primary IoMT nodes [14], we have eliminated the logistics costs, as well as the costs of the medical equipment, and significantly reduced the barrier to entry in the system. The approach is highly favourable, especially in the case of the elderly and those living in rural areas, as it is not possible to deploy high-class medical wearables in those regions.

3.3 Data Acquisition & Integration Layer

This layer handles the acquisition of data and enables its transmission. The User Interface is based on React and Vite, which gives it a highly responsive single-page application look and feel. This application has the functionality to operate in two modes to cater to various clinical scenarios. It can handle asynchronous video uploads for offline analysis and webcam streaming for real-time analysis.

To process the video frames in real-time, the frames are sent to the server through WebSockets. The data sent to the server is dynamically buffered to optimize the queries made to the server, which are used to process the video. The video processing routines are implemented with precise intervals. The intervals are fixed at 128 frames. At a standard video frame rate of 30 frames per second, the video duration of 4.27 seconds is sufficient to cover a complete cardiac cycle, which occurs 4 to 7 times in adults at rest.

3.4 Preprocessing Layer

The preprocessing layer is the most important purification component. In the raw video frames, a huge amount of irrelevant data is present, along with varying illumination pixels, which can easily dominate the cardiovascular signal. To obtain a clean physiological signal, a very strict preprocessing is applied. First, the highly optimized real-time face detection is performed using the MediaPipe library. Once the face is detected, the particular ROI of the face, which is usually focused on the highly vascularized regions such as the forehead and cheeks, is very strictly cropped to a standard size of 36x36 pixels, along with a numerical normalization in the range of [0, 1]. This minimizes the computation by reducing the data dimensionality to a greater extent.

Figure 2 shows visualization of the 36x36 ROI preprocessing pipeline. As can be observed from the low spatial resolution, the downsampling effectively preserves the macroscopic spatial topography of the face (Mean Frame) as well as the important pulsatile information (Difference) while acting as a natural filter against high-frequency environmental noise.

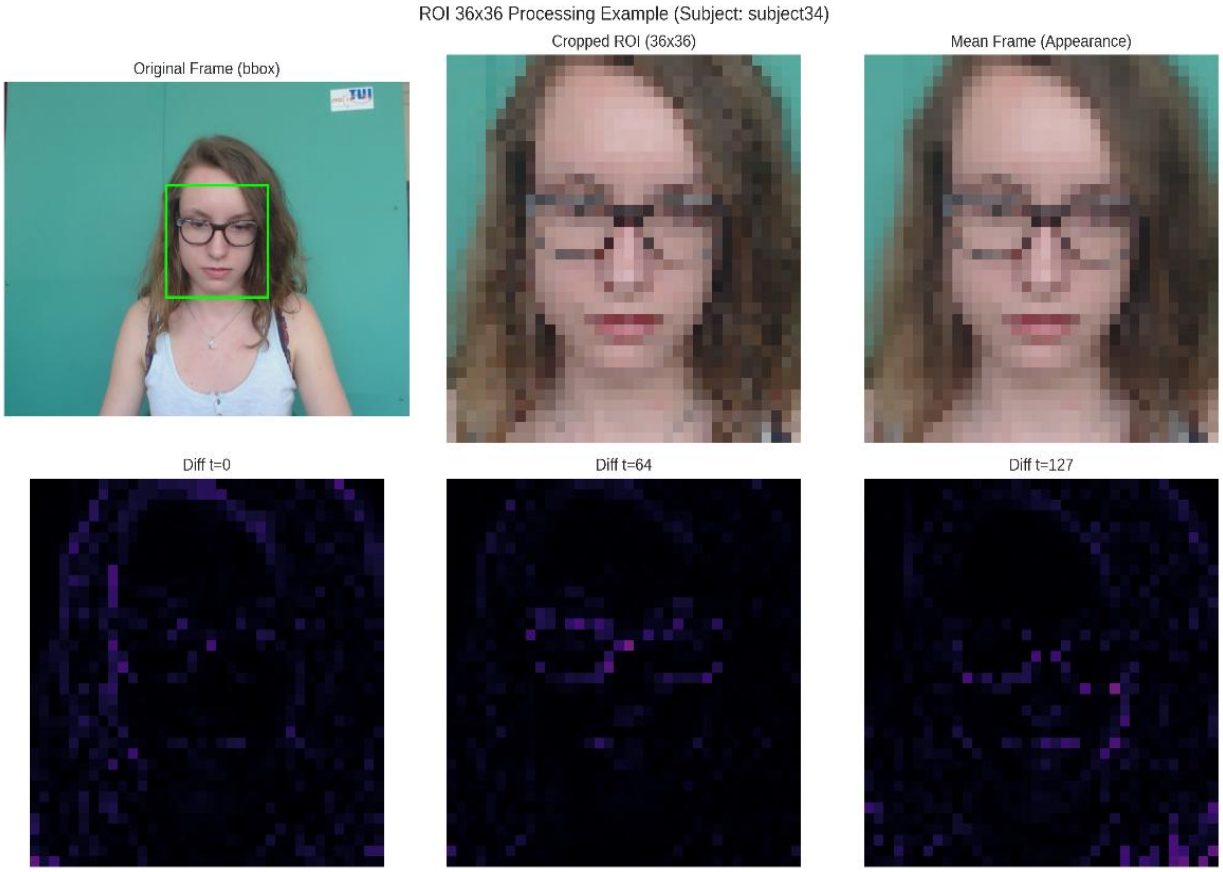

Figure 2. Visualization of the 36x36 ROI Preprocessing Pipeline

After conducting the sensitivity analysis on the input parameters, we eventually chose the 36x36 resolution. As visually represented in Figure 2, the 36x36 resolution effectively retains the macro spatial morphology of the face (Mean Frame). More pertinent is the fact that the 36x36 resolution retains the subtle dynamic colour changes required for effective pulse extraction (Difference). Increasing the resolution to 72x72 pixels or 128x128 pixels will lead to quadratic growth in the input pixel count. This, in turn, will result in an enormous growth in computational overhead (FLOPs), effectively acting as a bottleneck for real-time processing required for web deployment.

Similarly, the rationale behind choosing the 128-frame time window is based on physiological requirements and is not arbitrary. On the basis of a normal capture rate of 30 FPS, the time window of 128 frames equates to a period of 4.27 seconds. A normal resting HR for a human is between 60 and 100 Beats Per Minute (BPM), which equates to a frequency of 1 to 1.66 Hz. Therefore, this time window is adequate to capture 4 to 7 cycles of the HR. It is established

that this time window is adequate to capture periodicity in the signal such that the Fast Fourier Transform (FFT) is able to pick up the frequency peak. It is also established that if the time window is shorter than this, it is not sufficient to capture the required number of complete cycles to obtain accurate frequency, which is not possible. Furthermore, if the time window is longer, say 256 frames, it is equivalent to a delay time of nearly 9 seconds, which is not desirable.

The significant innovation under this category is the method adopted for the data bifurcation process, which is specific to the Two-Stream network. Rather than passing the raw sequence of frames to the network, the data bifurcation process specifically separates the static features from the dynamic features. First, the Mean Frame is computed by averaging all the pixel values of the entire sequence of frames. Further, the image acts as a filter that eliminates the dynamic features of movement while retaining the static features such as the skin tone, lighting patterns, and topology of the face. A set of Difference Frames are computed by subtracting each frame from the preceding one. This process of temporal differencing eliminates the static features of the background and the topology of the face while retaining the dynamic features arising out of the movement of the head and the colour changes of the BVP. Also, we built an active UI feedback loop to assess the quality of the input data. An oval shape alignment guide is rendered on the front end to ensure user compliance by maintaining the face at a stationary position and sufficiently illuminated at the optimum zone.

3.5 Edge/ Cloud Machine Learning (ML) Layer

The complete end-to-end workflow of the proposed spatial-temporal deep learning pipeline is illustrated in Figure 3.

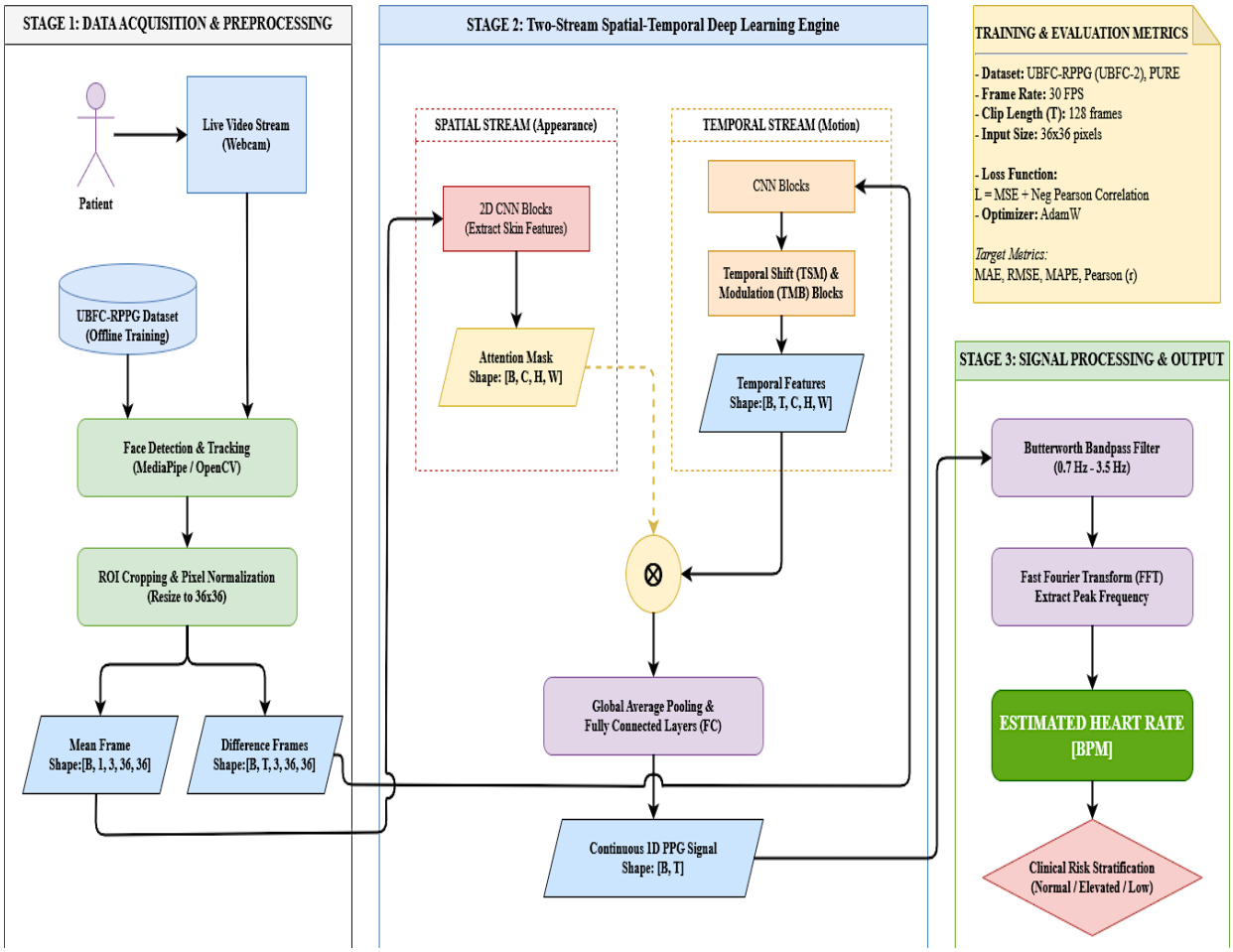

Figure 3. Proposed Spatial-Temporal Deep Learning Pipeline

After the data preprocessing step was performed, we applied a deep learning framework as the core predictive tool. We use this architecture because it is consistent with the branched nature of our pre-processed data. Moreover, this architecture optimizes the computational resources much better than the use of the traditional 3D CNN networks. In the Spatial Stream of the architecture, convolutional layers and the attention mechanism are applied to the 'Average Frame', from which the spatial masks are generated. These are actually a guide for the model to pay attention to the areas of the skin where the heartbeat signals are present. Meanwhile, the 'Difference Frames' will be inputted into the 'Temporal Stream' of the architecture. In the 'Temporal Stream,' the 'Difference Frames' will undergo the 'Temporal Shift Modules' and the 'Temporal Modulation Blocks'. It is possible for the network to clearly differentiate the changes in the skin pigmentation due to blood circulation and the actual movement of the head through the spatial masks.

However, this raw output signal is usually plagued with high-frequency noise due to the effects of camera quantization error and micro-expressions. In order to obtain a medical-grade signal that will be used to compute HRs, we apply a post-processing technique in signal processing is employed. First, the 1D signal is filtered using a Butterworth band-pass filter. This type of filter is defined by a range that is tightly constrained to range from 0.7 Hz to 3.5 Hz. This range is physiologically valid because it represents a range that includes the extreme limits of human HRs, which range from 42 to 210 BPM. After that, the signal is labelled to FFT processing. The FFT changes the nature of the signal to its frequency domain. In this case, the frequency component is obtained. This is the maximum component and is called the "dominant peak". This frequency is then multiplied by 60 to obtain the absolute HRs.

3.6 EHR & Clinician Dashboard Layer

Lastly, the computed metrics are automatically spreaded back to the React frontend, which functions as an interactive Clinical Decision Support System (CDSS) interface. This interface will be able to provide real-time physiological awareness in that it will show a numeric display of the estimated BPM in real-time, as well as a live plot of the expected PPG waveform. This is done in the form of a moving plot to give an understanding of the quality of the heartbeat.

The system also gives data in its raw form that is later converted into medical background. This is done by a process called risk stratification. The process is done by using a logic engine that always compares the computed BPM to medical criteria. The logic engine gives a classification of the status of the patient's heart health into a tier. The tier can be Normal status with HRs ranging from 60 to 100 BPM, an Elevated status with HRs higher than 100 BPM, and a Low status with HRs lower than 60 BPM. The visualization of the patient's status is an important part in the decision support in telehealth management.

3.7 Data Collection and Ground Truth Synchronization

For testing and training process, we have used the UBFC-RPPG (UBFC-2) dataset [15], which is a public dataset. This dataset has 42 different subjects. Each subject has two components. The first component is the face video without any compression, which has 30 frames per second. The second component is a .txt file that has the actual medical information of that subject. From this .txt file, we can obtain the actual Blood Volume Pulse (BVP) waveform and the actual HR value for each frame. When shooting videos for this dataset, each subject is connected to a pulse oximeter. This allows us to relate the subtle changes in colour that are detected on the face to the actual HR of that subject because this dataset is complete. In addition to the UBFC dataset [15], we also incorporated another dataset known as the PURE dataset [16] into our training and testing process. The data for this dataset was collected by researchers in a European lab environment. The dataset comprises 10 subjects aged between 18 and 35 years. The dataset comprises a total of 60 videos. Therefore, each of the subjects contributed around six video sequences. Just like the UBFC-RPPG dataset [15], the PURE dataset [16] also has accurate ground truth for the pulse waveforms and is recorded at 30 frames per second.

We also plan to test and train our model on other datasets, such as real-world clinical or telemedicine video data, to prove the applicability of our work. However, real-world video data of patients is very stringently regulated by laws regarding video data and cannot be publicly accessed for research purposes. Because of these stringencies for accessing data for our research, our current training and testing is limited to these two publicly accessible datasets. Moreover, to make this evaluation realistic, we have used a subject-independent 5-fold cross-validation technique. The total number of individuals is divided into five different groups. Our network is trained on four groups, and then we test our network on the remaining one.

One of the important engineering challenges in video data processing for physiological regression is to ensure absolute time alignment. The ground truth files contain raw BVP data and instantaneous HR data from a medical-grade pulse oximeter. The data ingestion pipeline processes video data by temporally segmenting the raw video sequences into distinct video clips. Each video clip contains T=128 frames. Based on the 30 FPS video recording rate, the video length is approximately 4.27 seconds. To allow the neural network to learn correct phase relationships, a video tensor with 128 frames is perfectly synchronized with exactly 128 data points in the ground truth BVP signal. Furthermore, to avoid the occurrence of severe I/O bottleneck effects in the epoch iteration process, all facial bounding boxes are pre-cropped in memory.

3.8 Model Training and Loss Formulation

Not like regular classification methods, rPPG analysis is basically a complex signal regression problem. The requirement is to accurately reconstruct both the amplitude and rhythm of the heart. Therefore, the research team chose the AdamW optimization algorithm along with a hybrid loss function. In particular, it uses a combination of MSE, which pushes the deviation in the amplitude to the lowest level, and Negative Pearson Correlation, which keeps the predicted wave in phase, close to the original signal structure. In a real clinical environment, it is vital to restore the morphology of the waves properly. Thanks to this, we can compute in-depth information, such as the Heart Rate Variability (HRV), which a noisy signal will surely ruin.

The loss function is a key part of the training. We combine two distinct metrics to form a Hybrid Loss. The first component is the Mean Squared Error (MSE) as shown in Equation (1), which penalizes the absolute amplitude differences between the predicted PPG waveform (y_{pred}) and the ground truth BVP (y_{gt}):

$$L_{MSE} = \frac{1}{T}\sum_{t=1}^{T}\left(y_{pred}(t) - y_{gt}(t)\right)^2 \tag{1}$$

While MSE is excellent for amplitude scaling, it is sensitive to baseline shifts and does not explicitly measure phase alignment. Therefore, the second component incorporates the Pearson Correlation Coefficient (r) as shown in Equation (2), which evaluates the linear correlation and structural similarity between the waveforms, independent of scale:

$$rr = \frac{\sum_{t=1}^{T}(y_{pred}(t) - \bar{y}_{pred})(y_{gt}(t) - \bar{y}_{gt})}{\sqrt{\sum_{t=1}^{T}(y_{pred}(t) - \bar{y}_{pred})^2}\sqrt{\sum_{t=1}^{T}(y_{gt}(t) - \bar{y}_{gt})^2}} \tag{2}$$

To optimize the network via gradient descent, we formulate the correlation as a minimization problem (Negative Pearson). The final Hybrid Loss (L_{total}) is defined in Equation (3) as:

$$L_{total} = \alpha \cdot L_{MSE} + \beta \cdot (1 - r) \tag{3}$$

where α and β are empirically determined balancing coefficients. By minimizing (r), the network strongly prioritizes the correct periodic rhythm of the cardiac cycle, which is mandatory for downstream applications like HRV analysis.

In order to train the model with the 36x36 ROI inputs, we chose the Adadelta optimization algorithm with an initial learning rate set to 0.05. In addition, in order to effectively control the training process, the StepLR scheduler is used to reduce the learning rate by a factor of 0.8 every 4 epochs. Furthermore, the training process is performed for 25 epochs with the batch size set to 32. As far as the hybrid loss function is concerned, equal weights are assigned to the terms by setting the weights α and β to 1.0. In addition, the temporal window length is set to $T = 10$, while the channel shift ratio is set to (η/T) = 0.25. Finally, the random seed is set to 20 in order to replicate the results.

3.9 Clinical Evaluation Metrics

In order to measure the effectiveness of the model, we decided to conduct a separate experiment on its chemical capabilities. Once we determined the value of the HR using a Faster Fourier Transform (FFT), we applied the following four indicators:

Mean Absolute Error (MAE): This is a measure of the clinical error margin. It is given in BPM. Let N be the number of video clips processed, the MAE is calculated as shown in Equation (4):

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\mathrm{HR}_{\mathrm{pred}}^{(\mathrm{i})} - \mathrm{HR}_{\mathrm{gt}}^{(\mathrm{i})}\right| \tag{4}$$

Root Mean Squared Error (RMSE): This is a measure of the stability of the system, where a larger error is penalized more heavily. This is a good indicator because a low error rate shows that the system is stable even in the presence of sudden and extreme motion artifacts, which would normally cause a system to fail catastrophically, as defined in Equation (5):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{HR}_{\mathrm{pred}}^{(\mathrm{i})} - \mathrm{HR}_{\mathrm{gt}}^{(\mathrm{i})}\right)^{2}} \tag{5}$$

Mean Absolute Percentage Error (MAPE): Since the accuracy of a HR deviation is clinically more important at some baseline than others, a 5 BPM deviation is more important at a baseline rate of 60 BPM than at a baseline rate of 150 BPM, the MAPE measures the system's performance based on the actual HR, as shown in Equation (6):

$$\mathrm{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\left|\frac{\mathrm{HR}_{\mathrm{pred}}^{(\mathrm{i})} - \mathrm{HR}_{\mathrm{gt}}^{(\mathrm{i})}}{\mathrm{HR}_{\mathrm{gt}}^{(\mathrm{i})}}\right| \tag{6}$$

Pearson Correlation Coefficient (p): Apart from metrics based on errors, we also measure the similarity between the predicted and reference physiological signals by using the Pearson correlation coefficient. The Pearson correlation coefficient is a measure used to calculate the linear correlation between the predicted rPPG signal and reference PPG signal. The Pearson correlation coefficient P close to 1 signifies a strong positive correlation as shown in Equation (7):

$$\mathrm{p} = \frac{\sum_{t=1}^{T}\left(\mathrm{PPG}_{\mathrm{pre}}(\mathrm{t}) - \overline{\mathrm{PPG}}_{\mathrm{pre}}\right)\left(\mathrm{PPG}_{\mathrm{gt}}(\mathrm{t}) - \overline{\mathrm{PPG}}_{\mathrm{gt}}\right)}{\sqrt{\sum_{t=1}^{T}\left(\mathrm{PPG}_{\mathrm{pre}}(\mathrm{t}) - \overline{\mathrm{PPG}}_{\mathrm{pre}}\right)^{2}\sum_{\mathrm{t}=1}^{\mathrm{T}}\left(\mathrm{PPG}_{\mathrm{gt}}(\mathrm{t}) - \overline{\mathrm{PPG}}_{\mathrm{gt}}\right)^{2}}} \tag{7}$$

where PPG_{gt} and PPG_{pre} are the ground truth and the predicted rPPG and $\bar{x}$ denotes the average operator. The spatially guided two-stream architecture always achieved highly competitive results in terms of both MAE and RMSE with our proposed phase-aware hybrid loss function.

3.10 ML for Remote Patient Management

While telemedicine today is limited to the diagnosis of a particular condition of a patient at a particular point of time, say, interpreting images of a patient through an X-ray to diagnose pneumonia, the term "Remote patient management" is a much broader term that handles the management of the physiological condition of a patient with the aim of personalized intervention, medication management, and timely detection of warning signs of a patient condition that might deteriorate suddenly. ML is a technology used for the management of the condition of a patient based on the large amount of data collected from patients, which has not been processed into useful information.

Normally, data has been segmented with limited data available include possible patients' symptoms through a chatbot interface or a physical logbook of daily blood pressure measurements. This, of course, was likely a number of lines with patients' symptoms typed out through a chatbot system or a physical logbook of daily blood pressure measurements. Yet, it was not until the advent of the boom in smart wearable technology and IoMT that this problem has been solved with the ability to collect sequences of data. In order to use this tremendous source of data, researchers have managed to use models especially suited for processing time sequences, like Recurrent Neural Networks (RNN) or LSTM. Theoretically, these algorithms perform extremely well at the task of sketching the 'baseline biological rhythm' for each individual, thereby detecting any abnormal signs early.

However, the use of physical hardware in monitoring has significant challenges in practical implementation. It is very impractical to ask elderly patients, for example, to wear smartwatches or attach body sensors for 24 hours a day. Most of these devices cause physical discomfort for the patient, which ultimately results in a decline in patient compliance

over a long period of time. The decline in patient compliance is known as sensor fatigue. Moreover, the deployment of medical wearables has significant limitations, such as the high cost of hardware devices and the need for battery maintenance.

To conceptualize the disruptive impact of our proposed contactless technology within this ecosystem, we augment the traditional taxonomy of ML in telemedicine. Table 1 explicitly situates the Two-Stream rPPG approach alongside established data modalities.

Table 1. Adding rPPG to Representative ML Techniques for Remote Diagnosis

Data Modality	**ML Technique**	**Diagnostic Application**	**Role in Telemedicine Contex**
Facial Video Sequences (rPPG)	Two-Stream CNN with Spatial-Temporal attention and TSM.	Contactless HR estimation, continuous vital sign tracking, arrhythmia detection.	Replaces physical contact sensors (wearables). Facilitates highly accessible, continuous remote monitoring requiring only a ubiquitous webcam. Ideal for tele-psychiatry, elderly care, and triage.
Physiologic al time-series (Wearables)	RNNs, LSTM models	Early detection of clinical deterioration.	Supports continuous remote diagnosis based on physical hardware distributed to patients.
Medical Images (Dermatology, X-ray)	CNNs	Skin lesion classification, radiological screening, diabetic retinopathy detection.	Enables automated image interpretation from episodic, patient-submitted static images, extending specialist diagnostic capabilities.

4. RESULTS AND DISCUSSIONS

4.1 Results

Regarding the evaluation of the results obtained, we designed and implemented a system using a 2-stream network to determine HR estimates from user-provided input data. HR estimation was performed under natural conditions. The prediction accuracy of the network is calculated by evaluating the estimated HR with respect to the ground truth, which is medical-grade, using a 5-fold cross-validation protocol.

To provide context for the performance comparison, we review recent advances in deep learning-based rPPG. In recent studies, several types of neural networks have been used in an attempt to improve smart monitoring devices used in clinical settings. Among those types, spatio-temporal [17] models and generative models, such as PulseGAN [18] have also been proposed to synthesize realistic pulse waveforms. Although wearable devices are commonly used today [19], camera-based PPG methods require solving many problems, specifically motion artifacts and skin tone variability issues [20]. In order to solve those problems, recent studies have introduced novel architectures such as transductive meta-learning [21], contrastive learning [22], and multi-scale networks [23].

Detailed comparison between the proposed model and other methods is shown in Table 2. As shown in the table, the suggested model has a highly competitive value in terms of MAE with a value of 0.75 bpm for UBFC-rPPG [15] and 0.73 bpm for PURE [16] datasets, while the RMSE is equal to 2.10 bpm for UBFC-rPPG [15] and 1.71 bpm for PURE [16] datasets. It should be noted that we decided to exclude the Pearson correlation coefficient (p) and MAPE from this evaluation. The reason is that the method of calculating the Pearson coefficient in previous documents is quite inconsistent , like in some places it is calculated directly on the rPPG signal, in others it is calculated on the HR value, making direct comparison difficult. In addition, the MAPE index is also rarely fully reported in baseline studies.

Therefore, to ensure objectivity and consistency, we only focus on two main measures, which are MAE and RMSE. A deeper analysis of the figures in Table 2 will be presented right in the next section.

Table 2. Comparison of the Accuracy for UBFC-rPPG [15] and PURE [16] Datasets

Method	**UBFC-rPPG**		**PURE**	
	MAE	RMSE	MAE	RMSE
PulseGAN (2021) [18]	1.19	2.10		
Meta-rPPG (2020) [21]	5.97	7.42		
Gideon (2021) [22]	3.70	4.90	2.1	2.6
AND-rPPG (2022) [24]	2.67	4.07		
STI (2022) [25]	3.88	6.23		
rPPG-FuseNet (2022) [26]	1.52	2.86		
MSSF (2022) [27]	0.74	1.87	1.121	2.419
CPulse (2023) [28]	1.06	1.48	0.98	1.94
REA (2023) [29]	*0.58*	**0.94**	1.23	2.01
STSC + MHFF (2023) [30]	2.15	3.82		
CMRPPGFormer (2024) [31]	0.59	1.77	*0.39*	*0.64*
ND-DeepRPPG (2024) [32]	**0.31**	*0.98*	**0.18**	**0.41**
LGI-rPPG-Net (2024) [33]	1.51	2.91		
Ours	0.75	2.10	0.73	1.71

(Note: Bold numbers represent the best performance and underlined, italic numbers represent the second best)

These results illustrate that machine slearning-based rPPG systems can be successfully built into intelligent telemedicine platforms. The integration of spatial attention and TSMs improves robustness to natural head movements, enabling more accurate HR estimation.

One of the major goals of this framework is to allow for the deployment of the system in a real-time environment over the web, avoiding the need for specific high-end hardware. In order to validate the claims made in this paper about the capabilities of the system in a real-time environment, the computational overhead of the system has been rigorously tested on a standard Central Processing Unit (CPU). An essential aspect of this testing is that it is done over the entire end-to-end processing of the system, from the first receipt of raw video frames to the final calculation of the HR. The experimental setup used the standardized temporal window of 128 frames. Using a standard capture rate of 30 frames/s, this equates to a time window of 4.27 seconds.

Table 3 shows The performance metrics of this end-to-end execution. As shown in Table 3, the deep learning model requires about 72.9 ms to process the entire 128-frame tensor. Most of the processing time is consumed by the pre-processing activities such as face detection and the calculation of the difference frames. The overall latency for the process is 451.2 ms. This indicates a processing speed of approximately 283 FPS. The processing speed of 283 FPS is quite high compared to the average of 30 FPS of a normal webcam connected to the internet. This shows that the proposed architecture is very practical in use. Once the camera is able to capture the initial video window of 4.27 seconds, the overall HR of the user is calculated by the web dashboard in less than 3 seconds. The overall memory consumption is also low for the proposed architecture, requiring approximately 337 MB of memory.

Table 3. End-to-End System Performance Benchmark (CPU-only, T=128 frames)

Metric	**Value**
Model Inference Time	72.9 ms
Total Pipeline Latency	451.2 ms
End-to-End Throughput	~283 FPS
Average CPU Usage	69.8 %
RAM Consumption	336.8 MB

4.2 Clinical Prototype and User Interface Integration

Besides quantitative evaluations, to bring the spatial-temporal rPPG model into practical application, we realized it is necessary to build a friendly and easy-to-operate user interface. Instead of just stopping at the level of running deep learning algorithms on a computer, we developed a web-based application acting as a CDSS. Currently, we chose this

approach because of its ability to run directly on normal web browsers without requiring the installation of additional specialized hardware.

As shown in Figure 4(a), the homepage interface highlights the feature of contactless vital sign measurement. At the same time, the system structure is designed with the criterion of data privacy placed on top (Privacy-by-Design). Although the real-time face video stream needs to be transmitted securely (through encryption protocols) to the inference server for processing by the deep learning model, the system ensures the analysis process happens instantly on temporary memory (in-memory processing). The sensitive video frames of the patient will be discarded immediately after extracting the heartbeat signal, completely not being stored on any database (Figure 4(b)). This is how we do it to maximally protect user privacy during the remote diagnosis process. The summary of key functionalities include hardware-independent, privacy-focused, and high-speed spatial-temporal inference.

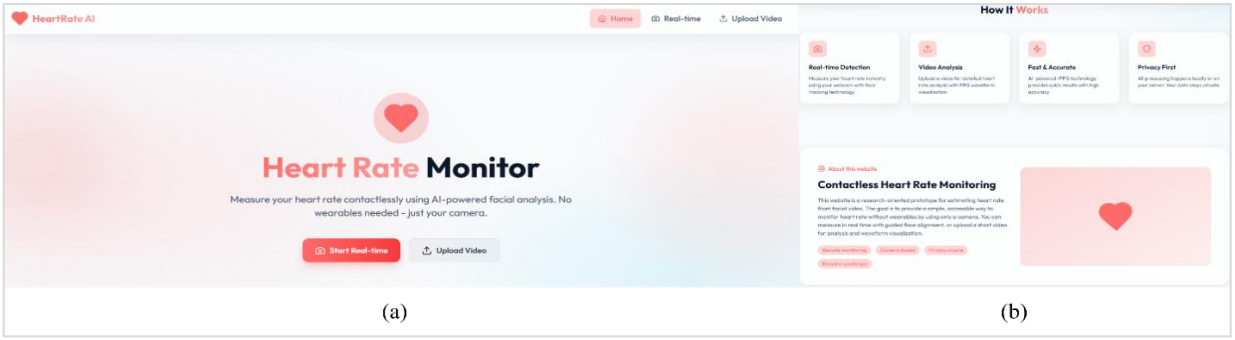

Figure 4. System Overview and Core Features (a) Interface of the eb-based rPPG CDSS (b) Summary of Key Functionalities

When measuring vital signs via camera, the biggest barrier is how to keep the user sitting in the correct posture, because just a twist or turn of the head will cause the Signal-to-Noise Ratio (SNR) to drop significantly. To overcome this problem right from the data collection stage, we have pre-designed a measurement scenario. Accordingly, the user is required to meet a few standard postures before the data is pushed into the predictive neural network (details are illustrated in Figure 5).

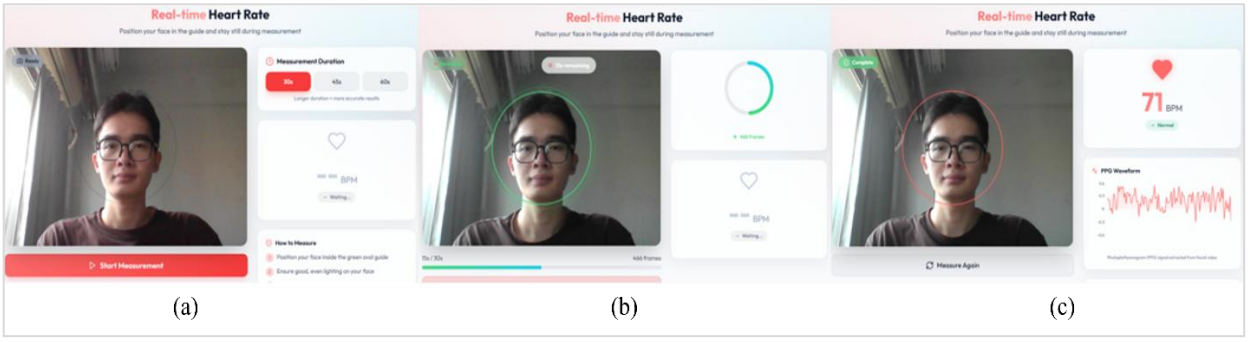

Figure 5 Design of the Experimental Workflows and Interfaces

Initially, the interface will require the user to choose the measurement duration with the following milestones: 30, 45, or 60 seconds (Figure 5a). The objective of this step in setting up is to ensure that the system is able to collect a certain number of heart cycles. This will ensure that the signal analysis in frequency domain occurs in a stable manner. The idea in this step is to ensure that the face alignment mechanism is actively integrated into the screen. For instance, a guide frame will be provided to assist in aligning the face in the required coordinates. The face will always be tracked in the process, while at the same time visual cues are provided to the user. Once the required distance and lighting are in place at optimal levels, the frame turns green instantly, and the progress bar begins to move (Figure 5b). This step serve as a physical filter for noise right from the outer ring of the circle.

When the measurement time ends, the results are immediately putted into the medical monitoring interface (Figure 5c). The difference is that instead of only displaying predicted numbers, the system will redraw the entire continuous PPG waveform. Thanks to this, remote doctors can still visually observe the HR morphology of the patient. At the same time, according to the risk mentioned above in the previous section, the software would automatically use the measured value (for example, 71 BPM) to compare and determine the health status (for example, "Normal").

In Figure 5, the design of the experimental workflows and interfaces includes a calibration setup for temporal window selection (a), a live recording interface that utilizes bounding ovals and real-time progress indicators to reduce motion noises (b), and a clinical dashboard showcasing the estimated HR, automated risk stratification, and the reconstructed PPG waveform (c).

4.3 Ablation Study

In order to assess the contribution of each of the components in the framework, we carried out an ablation study on the Temporal Stream of the proposed framework on the PURE dataset [16]. We tested three different variants of the Temporal Stream: (1) the original In-Place Temporal Shift Module (TSM), (2) the Residual TSM, and (3) the Residual TSM w/o Temporal Attention Module. The experiments were carried out for three different inference window sizes: W = 10s, 20s, 30s. The results are presented in Table 4.

Table 4. Ablation Study of the Temporal Stream on the PURE Dataset [16]

Method	**W = 10s**		**W = 20s**		**W = 30s**		
	MAE ↓	**RMSE ↓**	**MAE ↓**	**RMSE ↓**	**MAE ↓**	**RMSE ↓**	**p ↑**
In-place TSM	0.937	2.489	**0.738**	**1.789**	0.915	1.706	**0.677**
Residual TSM	**0.820**	**2.283**	0.772	1.867	**0.887**	**1.691**	0.674
Residual TSM w/o TAM	0.997	3.569	0.858	2.088	1.025	2.099	0.667

The data also clearly illustrates the importance of the Temporal Attention Module. In our experiments, when we took away the TAM, error rates shot up on all window sizes. For example, when W = 30s, MAE increased from 0.887 to 1.025 BPM, and RMSE increased from 1.691 to 2.099 BPM. This proves that temporal attention is indeed a critical component because it will automatically assign more weights to frames where strong physiological signals exist and, at the same time, suppress frames where strong signals are corrupted by sudden head movements.

Additionally, based on the comparison results of the two methods of shifting, a performance trade-off is also identified. Although the In-place TSM achieves a successful performance at a time window W = 20s, the Residual TSM achieves a better performance at time intervals W = 10s and W = 30s. The reason is in the Residual TSM, skip connections have been applied to prevent the loss of critical spatial information in the channel shifting procedure. Moreover, this study also successfully develops a telemedicine system that operates on the web, bridging the gap between deep learning theory and its practical application.

4.4 Discussion

In general, even though the web-based rPPG system has a lot of potential in a clinical environment, it is a completely different issue when we attempt to transfer it from a laboratory environment to real-world application. When we attempt to do that, we actually encounter a few challenges in terms of technology and ethics. The first problem of this study is still having to use a public dataset instead of having actual telehealth data.

Furthermore, environmental noises are one of the fundamental challenges faced by all optical camera-based systems. Although the spatial-temporal architectures are successful in reducing noises during natural conversational movements, they are highly susceptible to extreme movements. Any sudden head rotations or rigorous physical activities can cause non-linear pixel distortions almost instantaneously. It is in these situations that it is extremely difficult for the network to distinguish the actual HR signal from the motion artifacts. Another trouble is lighting. For example, if a user sits in a dark room watching TV, the continuously flickering screen will create fake frequencies in the video. If these frequencies coincide exactly with the human HR band (from 0.7 to 3.5 Hz), it will make the system predict incorrectly. Not to mention, the improper sitting posture of the patient or the inconsistent quality of personal

camera devices also makes the model unstable. The collected data is both noisy and prone to generating false alarms, or worse, missing the disease. In this study, we did not control empirical experiments to measure performance under specific motion and lighting conditions. In this first stage of proof of concept, the main focus is on the design of the system architecture and proving the basic feasibility of the architecture on a benchmark. A detailed evaluation of the environmental noise variables is not within the baseline. As future work, it is planned to develop particular data sets with controlled variables to evaluate the performance with particular lighting conditions and movement speeds.

Regarding the ethical aspect, there is a problem in the training process of the rPPG predictive model, which is the fairness of the algorithm regarding skin colour. The nature of rPPG is measuring light penetrating the skin and then bouncing back from the blood vessels. The problem lies in the fact that dark-skinned people (Fitzpatrick skin types IV-VI) have a lot of Melanin which absorbs almost all visible light. In this research, we have the limitations of our training and validation data in terms of ethnic diversity. Our system is currently trained and tested using only two data sets: UBFC-RPPG [15] and PURE [16]. A visual assessment of the 42 subjects in the data set provided by UBFC-RPPG [15] showed a lack of diversity in terms of skin tones, as all subjects were predominantly lighter in skin tone (Caucasian). The PURE dataset [16] had around 10 subjects in a European laboratory setting and lacked subjects with darker skin tones. Since there is a lack of quantitative data in these data sets in terms of skin tones across the range of the Fitzpatrick skin type scale (IV-VI), we were not able to subgroup these data and quantify the bias in skin tones. As a result, we were able to identify a lack of diversity in these data as a systemic problem in our research. The skin tone problem in our system in this research is a systemic problem that needs urgent solution in future research in rPPG by having more diverse physiological data.

Therefore, the pulse signal obtained (SNR) from them is extremely weak. The majority of current deep learning models are entirely trained using data from light-skinned people. As a result, when taken to be tested on ethnic minority groups, the error rate skyrockets. In healthcare, being biased like this is very dangerous. If the system is used on a mass scale, and the machine makes biased risk diagnoses simply because of the patient's skin colour, what the consequences would be. Therefore, future rPPG studies are required to collect much more diverse data. At the same time, programmers must manually design additional specialized loss functions to force the model to be less biased; only then can medical technology truly be fair.

Not just stopping at the issue of algorithmic bias, we also have to face other major barriers regarding privacy, security, and the explainability of the model. In the remote healthcare environment, data is scattered everywhere, so the risk of cyberattacks targeting IoMT devices, home Wi-Fi networks, or cloud systems is very high. And unless we are able to encrypt this data appropriately, it is not easy to convince patients to share their sensitive health data with us.

From a system perspective, the current architecture is operating by pushing each video frame up to the cloud server (via FastAPI) for computation. This method truly exposes too many drawbacks: high latency, consuming bandwidth (especially in remote areas), and especially the risk of violating strict security standards like HIPAA or GDPR. Therefore, the direction we intend to deploy is building an app on mobile phones. To be able to run on phones without making the device hot or draining the battery quickly, we must compress the model size down. The solution is to replace the heavy CNN layers with a lighter network like MobileNetV3. After that, combining weight pruning techniques and Post-Training Quantization (PTQ) to compress the weights from 32-bit floating-point format down to 8-bit integers. Finally, we intend to convert the model into TensorFlow Lite format to run offline locally. By doing this, we can both guarantee the issue of not leaking data onto the network and also ensure an extremely smooth real-time response speed. In the long run, remote healthcare will also combine federated learning to train the model without needing to gather user data into one place or integrate digital twins. But no matter how far technology advances, it still requires doctors, engineers, and regulatory agencies to sit down together to establish common standards, ensuring AI serves patients in the safest and fairest way possible.

4.5 Case Studies And Applications

The process of transferring rPPG technology from the laboratory environment to practical application has proven great values in terms of both clinical and operational aspects. Specifically, through integrating the spatial-temporal deep learning network onto common web platforms, we can completely turn videos from standard cameras into a proactive diagnostic tool. Practical trials confirm that this solution helps enhance accuracy during diagnosis, improve patient compliance, and boost the response speed of the entire system. The practical impact of this technology is most clearly demonstrated through the three key medical areas below.

First is the field of telecardiology and chronic disease management. Cardiovascular Disease (CVD) to date remains the highest cause of death worldwide. Effectively controlling this group of diseases strictly requires continuous monitoring throughout a long period, instead of just relying on sparse periodic check-ups. However, the drawback of current ambulatory medical monitoring devices (such as Holter ECG monitors or blood pressure cuffs) is their bulkiness, more or less causing disruption to the daily life of the patient. Therefore, the web-based rPPG solution allows users to self-check their vital signs much more simply because they only need to sit in front of a phone or computer camera. At this time, the deep learning model will process the video data to outline a physiological baseline dedicated specifically to each individual. Thanks to that, with just a slight fluctuation such as an increased resting HR or an irregular interval between beats, doctors will immediately notice. Through automatically generated clinical reports, the system helps warn early of the risk of health deterioration.

The second thing to mention is the field of mental health assessment and stress level measurement. This is an extremely huge advantage of rPPG but is often forgotten in remote healthcare platforms. Our psychological states, for example when being heavily stressed or clinically depressed, actually relate very closely to the Autonomic Nervous System (ANS). Fortunately, modern deep learning models already have the ability to restore the PPG waveform in a continuous and extremely robust manner. From this waveform, we can extract the HRV. Understood simply, HRV measures the physiological change in terms of time between consecutive beats, and it is considered by the medical community as an accurate non-invasive biomarker to assess the balance state of the ANS. This physiological parameter plays a role as a vital physiological marker in the monitoring of physiological stress through video consultation sessions. This brings very huge value, because it provides psychologists with objective physiological numbers to cross-reference with their behavioural observations, thereby offering treatment regimens sticking closer to actual data.

Third is the story of elderly care and creating Ambient Assisted Living (AAL). Any doctor who has ever worked in an elderly care environment will understand, making patients comply with wearing long-term tracking devices is an extremely huge barrier. Elderly people, especially those with cognitive decline due to Alzheimer's disease or dementia, usually operate very poorly with smart wearable devices (wearables). Constantly having to remember to charge the battery every day, the interface being complicated, plus the cumbersome feeling on the body gradually generates a condition that the clinical community often calls "sensor fatigue". At this time, the contactless rPPG system embedded directly into smart home screens, televisions or tablets becomes a perfect alternative solution. Thanks to being well optimized, spatio-temporal neural networks have more than enough capacity to run smoothly on standard edge devices without causing any annoyance. The system will just need to monitor the high-risk group right while they are doing their normal activities such as eating or watching TV. This will not only help to detect dangerous cardiovascular events right away, but it will also maintain autonomy and comfort for the patient.

5. CONCLUSION

Remote health care is slowly moving away from passive examination and towards active and continuous monitoring. In this direction, our study was able to successfully develop a contactless HR measurement system directly in a web browser with its foundation in a Two-Stream Spatial-Temporal Architecture. The capability to independently extract spatial features and time transformations in facial video data allows the presented model of rPPG to use regular webcams to monitor vital signs with high reliability. The data collected in the experiments shows that this system is capable not only of extracting HR signals with high accuracy but also has the potential to act as a CDSS on its own. This opens doors to the potential benefits that can be seen in online cardiovascular examinations, mental health examinations, elderly health monitoring, and so on without the patient having to wear cumbersome devices.

However, there are still some challenges for the way from the lab to the real-world environment of rPPG technology. The biggest challenges are the degradation of performance caused by environmental noise, model bias with different skin tones, and strict medical data privacy policies. In order to address these challenges, the research team has a strong belief that the best solution is to transform from cloud inference to Edge AI with the help of compression techniques. This not only ensures high performance but also solves the issue of personal data leakage comprehensively. Besides, for contactless technology to win more trust from users, transparency of the algorithm is essential. Therefore, Federated Learning or XAI will be the focus of our next research, targeting an intelligent, proactive, and patient-centric medical ecosystem.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for the suggestions to improve the paper.

FUNDING STATEMENT

The authors received no funding from any party for the research and publication of this article.

AUTHOR CONTRIBUTIONS

Do Duc Thinh: Software, Conceptualization, Methodology, Validation, Writing – Review & Editing;
Nguyen Duc Manh: Conceptualization, Methodology, Validation, Writing – Review & Editing;
Nguyen Thi Bich Ngoc: Validation, Writing – Review & Editing;
Vu Van Huan: Methodology, Validation, Writing – Review & Editing.

CONFLICT OF INTERESTS

No conflict of interests were disclosed.

ETHICS STATEMENTS

Our publication ethics follow The Committee of Publication Ethics (COPE) guideline. https://publicationethics.org/

DATA AVAILABILITY

The data is publicly available. This research did not require IRB approval because it used publicly available secondary data with no identifying personal information.

REFERENCES

[1] P. C. Ahanotu, D. A. Adedigba, R. Hasan, and S. Palaniappan, "Deep learning-based automatic detection and diagnosis of tuberculosis from chest X-ray images: A comprehensive analysis," *Journal of Informatics and Web Engineering*, vol. 5, no. 1, pp. 69–85, Feb. 2026, doi: 10.33093/jiwe.2026.5.1.5.

[2] S. Palaniappan, R. Logeswaran, K. Subaramaniam, O. Baker, and B. N. Dung, "Training the brain: A machine learning approach to predicting wellbeing through intentional thought pattern modification," *Journal of Informatics and Web Engineering*, vol. 4, no. 3, pp. 64–89, Oct. 2025, doi: 10.33093/jiwe.2025.4.3.4.

[3] M. Rossi, and S. Rehman, "Integrating artificial intelligence into telemedicine: Evidence, challenges, and future directions," *Cureus*, vol. 17, no. 8, 2025, doi: 10.7759/cureus.90829.

[4] U. Chaturvedi, S. B. Chauhan, and I. Singh, "The impact of artificial intelligence on remote healthcare: Enhancing patient engagement connectivity and overcoming challenges," *Intelligent Pharmacy*, 2025, doi: 10.1016/j.ipha.2024.12.003.

[5] W. Wang, A. C. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote PPG," in *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479-1491, July 2017, doi: 10.1109/TBME.2016.2609282.

[6] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, pp. R1–R39, 2007, doi: 10.1088/0967-3334/28/3/r01.

[7] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics Express*, vol. 16, no. 26, pp. 21434, 2008, doi: 10.1364/oe.16.021434.

[8] M. Poh, D. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Express*, vol. 18, no. 10, pp. 10762, 2010, doi: 10.1364/oe.18.010762.

[9] G. de Haan, and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," in *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878-2886, Oct. 2013, doi: 10.1109/TBME.2013.2266196.

[10] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote Heart Rate Measurement From Highly Compressed Facial Videos: An end-to-end deep learning solution with video enhancement," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 151–160, 2019, doi: 10.1109/iccv.2019.00024.

[11] Z. Yu et al., "PhysFormer++: Facial video-based physiological measurement with slowfast temporal difference transformer," *International Journal of Computer Vision*, vol. 131, no. 6, pp. 1307–1330, 2023, doi: 10.1007/s11263-023-01758-1.

[12] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," *Lecture Notes in Computer Science*, pp. 356–373, 2018, doi: 10.1007/978-3-030-01216-8_22.

[13] W. -N. Lie, D. Q. Le, P. -H. Huang, G. -H. Fu, A. Nguyen Thi Quynh, and Q. Nguyen Quang Nhu, "A two-stream deep-learning network for heart rate estimation from facial image sequence," in *IEEE Sensors Journal*, vol. 24, no. 24, pp. 42343-42351, Dec.15, 2024, doi: 10.1109/JSEN.2024.3483629.

[14] V. K. Damera, R. Cheripelli, N. Putta, G. Sirisha, and D. Kalavala, "Enhancing remote patient monitoring with AI driven IoMT and cloud computing technologies," *Scientific Reports*, vol. 15, no. 1, p. 24088, 2025, doi: 10.1038/s41598-025-09727-z.

[15] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, "Unsupervised skin tissue segmentation for remote photoplethysmography," *Pattern Recognition Letters*, vol. 124, pp. 82–90, 2019, doi: 10.1016/j.patrec.2017.10.017.

[16] R. Stricker, S. Muller, and H. -M Gross, "Non-contact video-based pulse rate measurement on a mobile service robot," *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1056–1062, 2014, doi: 10.1109/roman.2014.6926392.

[17] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *Proceedings of the British Machine Vision Conference (BMVC),* 2019.

[18] R. Song, H. Chen, J. Cheng, C. Li, Y. Liu, and X. Chen, "PulseGAN: Learning to generate realistic pulse waveforms in remote photoplethysmography," in *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 5, pp. 1373-1384, May 2021, doi: 10.1109/JBHI.2021.3051176.

[19] S. Majumder, T. Mondal, and M. J. Deen, "Wearable Sensors for Remote Health Monitoring," *Sensors*, vol. 17, no. 1, pp. 130, 2017, doi: 10.3390/s17010130.

[20] E. M. Nowara *et al*., "Towards a deeper understanding of skin tone and motion in remote photoplethysmography," *CVPR Workshops*, 2021.

[21] E. Lee, E. Chen, and C.-Y. Lee, "Meta-rPPG: Remote heart rate estimation using a transductive meta-learner," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[22] J. Gideon, and S. Stent, "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)*, 2021.

[23] H. Yu, H. Lee, and K. Sohn, "Multi-scale spatio-temporal feature learning for remote photoplethysmography," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022.

[24] Z. Chen, W. Wang, and A. C. Bovik, "AND-rPPG: Adaptive Noise Decoupling for Robust Remote Photoplethysmography," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2022.

[25] Y. Niu, H. Luo, and W. Wang, "STI: Spatio-temporal interaction network for remote photoplethysmography estimation," *IEEE Transactions on Instrumentation and Measurement*, 2022.

[26] J. Cheng, H. Chen, and X. Chen, "rPPG-FuseNet: Remote photoplethysmography estimation via multi-feature fusion network," *IEEE Access*, 2022.

[27] C. Zhao, M. Zhou, W. Han, and Y. Feng, "Anti-motion remote measurement of heart rate based on region proposal generation and multi-scale ROI fusion," in *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1-13, Art no. 5012413, 2022, doi: 10.1109/TIM.2022.3169567.

[28] Z. Zhang, H. Wang, and L. Yin, "CPulse: Camera-based pulse estimation via deep learning," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[29] H. Li, X. Chen, and Y. Wang, "REA: Robust efficient attention network for remote photoplethysmography," *Biomedical Signal Processing and Control*, 2023.

[30] X. Liu, H. Wang, and Z. Yu, "STSC + MHFF: Spatio-temporal signal compensation with multi-head feature fusion for rPPG estimation," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, 2023.

[31] X. Ma, Z. Wang, X. Liu, and H. Kuang, "CMRPPGFormer: 3-D Spatio-temporal convolutional modulation transformer network for remote heart rate estimation," in *IEEE Sensors Journal*, vol. 24, no. 19, pp. 30275-30286, Oct.1, 2024, doi: 10.1109/JSEN.2024.3407816.

[32] S. -Q. Liu, and P. C. Yuen, "Robust remote photoplethysmography estimation with environmental noise disentanglement," in *IEEE Transactions on Image Processing*, vol. 33, pp. 27-41, 2024, doi: 10.1109/TIP.2023.3330108.

[33] M. Amin *et al*., "LGI-rPPG-Net: A Lightweight encoder–decoder network for remote photoplethysmography estimation," *Biomedical Signal Processing and Control*, 2024.

BIOGRAPHIES OF AUTHORS

	**Do Duc Thinh** is currently a Research Intern at the National Chung Cheng University (CCU), Taiwan, under the TEEP Program. He is a final-year undergraduate student majoring in Information Technology at the University of Transport and Communications, Vietnam. His research interests include Computer Vision, Deep Learning, and their applications in healthcare and smart systems. His recent work focuses on 3D human skeleton extraction and medical image analysis using deep learning techniques. He can be contacted at thinh212610443@lms.utc.edu.vn.
	**Nguyen Duc Manh** is a third-year undergraduate student majoring in Information Technology at the University of Transport and Communications, Vietnam. His research interests include Deep Learning, Computer Vision, Large Language Models (LLMs), Vision-Language Models (VLMs), and intelligent systems for healthcare applications. His recent work focuses on medical image analysis and physiological signal estimation using deep learning techniques. He can be contacted at manh232631031@lms.utc.edu.vn.
	**Nguyen Thi Bich Ngoc** received Bachelor and Master in Information Technology at University of Transport and Communications in 2012 and 2014, respectively. She is currently a lecturer at Sao Do University. Her research interests include network administrator, software engineering, and image processing. She can be contacted at nguyenbichngoc1990@gmail.com.
	**Vu Van Huan** received Bachelor at Hai Phong University of Management and Technology and master at Vietnam National University in 2005 and 2011, respectively. He is currently a lecturer at Hanoi University of Natural Resources and Environment, Vietnam. His research interests include machine learning, biomedical engineering, software engineering. He can be contacted at vvhuan@hunre.edu.vn.