FOREX Publication
Open Access | Rapid and quality publishing

International Journal of Electrical and Electronics Research (IJEER)
Research Article | Volume 13, Issue 4 | Pages 837-851| e-ISSN: 2347-470X

A Deep Learning-Based Approach for Heart Rate Monitoring through Combined Convolutional and Generative Networks Using Facial Videos

Jyostna J[1]*, and Satyanarayana Penke[2]

[1]Research Scholar, Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh-522502, India

[1]Department of Electronics and Communication Engineering, CVR College of Engineering, Vastunagar, Ibrahimpatnam (M), Hyderabad, Telangana-501510, India

[2]Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Andhra Pradesh-522502, India

*****Correspondence:** *Jyostna J; Email: jyostnajulakanti@gmail.com*

ABSTRACT- This research presents a deep learning-based architecture that uses facial video-extracted remote Photoplethysmography (rPPG) to non-invasively estimate heart rates. The proposed system addresses limitations in signal fidelity and scalability by integrating a Conditional Generative Adversarial Network (CGAN) to enhance the quality of raw rPPG waveforms and a 1D Convolutional Neural Network (CNN) for regression-based prediction of heart rate in beats per minute (BPM). Unlike traditional single-stream models, our framework supports concurrent processing of facial video streams, improving computational efficiency and applicability in real-time, multi-subject environments. Video data is pre-processed through facial Region of Interest (ROI) detection, spatial averaging in alternative colour spaces (YUV/LAB), and temporal filtering before being subjected to CGAN-driven denoising. A mean absolute error (MAE) of 2.3 BPM, accuracy of 95% and a Pearson Correlation Coefficient (PCC) of 0.92 versus reference signals were achieved by the CNN regressor when trained on enhanced signals according to the UBFC-rPPG dataset. Experimental results demonstrate the robustness of the developed model to lighting variation, head motion, and skin tone diversity. The proposed pipeline is well-suited for applications in telemedicine, contactless fitness monitoring, and smart surveillance systems requiring real-time physiological assessment. Real-time video streams have been used to test the suggested model, which shows little variation between the ground truth and the actual heart rate values. This low prediction error demonstrates the model's resilience and appropriateness for applications involving real-time physiological monitoring.

Keywords: Remote Photoplethysmography (rPPG), Conditional Generative Adversarial Network (CGAN), Video Processing, Convolutional Neural Networks.

ARTICLE INFORMATION

Author(s): Jyostna J, Satyanarayana Penke;

Received: 16/07/2025; **Accepted:** 08/12/2025; **Published:** 25/12/2025;
E- ISSN: 2347-470X;
Paper Id: IJEER 250716-10;
Citation: 10.37391/ijeer.130424
Webpage-link:
https://ijeer.forexjournal.co.in/archive/volume-13/ijeer-130424.html

Publisher's Note: FOREX Publication stays neutral with regard to jurisdictional claims in Published maps and institutional affiliations.

1. INTRODUCTION

The possibility of measuring physiological phenomena like heart rate in a non-invasive and unobtrusive way has come to the forefront in the past years, particularly in the field of telemedicine, mobile health, and smart environments. The essential biomarker is Heart Rate (HR), which is related to exercising, stress, feelings, and the condition of the heart. However, traditional techniques of HR metrics include contact technologies, including electrocardiography (ECG), PPG, and pulse oximetry [1-3]. Although they are correct, such strategies demand physical sensor connection to the body, which may cause pain during prolonged monitoring and restrict the number of settings they can be used in, especially in remote or mass-coordinated ones.

To overcome these limitations, researchers have come up with a method called contactless, where instead of touching the skin, a rPPG provides the measurement of the pulse rate acquired by analyzing the slight, but noticeable, colour changes on the skin caused by variations in blood volume in a regular RGB camera [4, 5]. In some cases, the rPPG signal is obtained by taking a sample of the face, forehead or the cheeks and the periodicity of the waveform of the signal is then used to estimate beats per minute (BPM). Although rPPG signal acquisition has good potential, it is fundamentally vulnerable to environmental noise, light changes, movements of the head, and low signal-to-noise ratio (SNR), decreasing the accuracy of the measurements [6].

The recent progress in deep learning has allowed building more robust estimation pipelines with the help of extracting discriminative features, as well as learning complex dynamics in a video stream. The ability of Convolutional Neural Networks (CNNs) to produce localized features with sequential inputs has led to their application in spatial-temporal analysis of rPPG signals [7, 8]. However, CNN-based models alone might not succeed in inhibiting noise or restoring answerable physiological signals under uncontrolled circumstances.

In order to solve this, Conditional GANs (CGANs) have been applied to rPPG waveform improvement through learning to map approximately noisy input signals to their de-noised representations conditioned on spatial context [9]. Such models have demonstrated superior performance in generating physiologically plausible rPPG signals, which, when fed into CNN regressors, result in improved heart rate estimation.

This research provides a hybrid deep learning architecture in this study that combines a Conditional Generative Adversarial Networks (CGAN) for rPPG signal augmentation with a one-dimensional convolutional neural network (CNN) for BPM prediction. The suggested system can be used in practice in the monitoring of ICUs, sports analytics, and driver fatigue prevention. The training and validation of the model are conducted on the UBFC-rPPG dataset [10]. Here are the main points of this work:

- Design a CGAN-enhanced pipeline for robust rPPG signal refinement under variable lighting and motion.
- Design a CNN-based regressor that learns heart rate patterns from denoised rPPG
- To confirm the validity of the performance and generalization ability of the model, perform extensive experiments on a benchmark dataset.

The structure of the paper is as follows: *Section 2* describes a concise overview of existing research on both contact-based and contactless heart rate estimation techniques. *Section 3* details the proposed methodology. *Section 4* describes the dataset and outlines the experimental setup. And the results and their analysis. Finally, *section 5* concludes the paper.

2. LITERATURE REVIEW

The domain of remote photoplethysmography (rPPG) has gained considerable traction due to its ability to perform non-contact physiological monitoring. By analyzing subtle skin color variations induced by cardiac pulse, rPPG eliminates the need for physical sensors. Prior research on measuring face rPPG has mostly used conventional signal processing techniques to examine tiny colour changes on facial areas of interest (ROI) [11, 12,13,14] Such an innovation will make possible novel applications in telemedicine, fitness tracking, and affective computing that the existing contact-based technologies, such as ECG and PPG, cannot serve because of hardware limitations and the intolerability by user [15].

Due to increased demand for remote health monitoring, deep learning methodologies have been developed to increase the accuracy and real-time capability of rPPG-based heart rate estimation systems. This survey involves classical approaches to detecting the heart rate, the development of rPPG, modern deep networks, and multi-video processing advances.

2.1. Conventional Heart Rate Monitoring Techniques

Electrocardiography (ECG): Implants electrodes that perceive electrical impulses in the heart. Although truthful, ECG setups tend to be big and obtrusive [1].

Photoplethysmography (PPG): Detects changes in blood volume through optical sensors and LEDs (infrared or green) positioned on the skin surface [2].

Pulse Oximetry: Although primarily used during oxygen standby and displaying heart rate, the device monitors light uptake via the ear or finger [3].

But these contact-based techniques have limitations, use incurs discomfort during all time use, sensitivity to motion artifacts, and cannot be used in large-scale or remote applications in an easy manner [16].

2.2. Remote Photoplethysmography (rPPG)

rPPG encourages ordinary RGB cameras to record video of the face and identify minute variations in the skin color that happen as a result of blood movement. Such variations are then converted to signals on heart rate. The main steps in the estimation by rPPG will be the following:

Face Detection and ROI Selection: In a conventional process, the forehead or cheek area is chosen as an area of steady signal acquisition [17].

Color Space Conversion: To more easily isolate chrominance variations, RGB signals are likely to be converted to YUV, HSV or LAB [5].

Temporal Filtering: Noise unrelated to the physiological frequency band is attenuated or rejected by the use of bandpass filters (0.7–4 Hz) [18].

BPM Estimation: Fast Fourier Transform (FFT), Wavelet Transform, or machine learning regressors predict the heart rate [19].

Although rPPG is a cost-efficient and contactless technique suitable for large-scale use, it is limited by a weak signal-to-noise ratio and sensitivity to head movement and variations in illumination.

2.3. Deep Learning Models in rPPG Estimation

To overcome the limitations of classical signal processing, several end-to-end deep learning approaches have been introduced [20-22], which directly estimate rPPG signals and other physiological parameters from facial video frames as input.

2.3.1. Convolutional Neural Networks (CNNs)

CNNs extract temporal-spatial features from video frames. Models such as ResNet, VGG, and DenseNet have been adapted for rPPG tasks [23]. They perform well under stable conditions but are less effective under dynamic lighting or motion artifacts.

2.3.2. Recurrent Neural Networks (RNNs) and LSTMs

Recurrent networks capture temporal dependencies between consecutive frames. LSTM-based models [24-26] have achieved improved performance on datasets like PURE and UBFC, although they are computationally intensive and vulnerable to frame dropout.

2.3.3. Generative Adversarial Networks (GANs)

GANs are used to denoise rPPG signals. Conditional GANs (CGANs), in particular, generate synthetic clean waveforms conditioned on noisy input [9]. Chen et al. proposed a GAN-assisted model for generating high-fidelity photo plethysmographic signals, significantly improving BPM accuracy [8].

2.3.4. Hybrid Models

Combining CNNs for feature extraction and GANs for signal refinement has yielded state-of-the-art results [8].

Literature reveals a consistent evolution in rPPG-based heart rate estimation, from traditional contact sensors to sophisticated, deep learning-powered, contactless systems. Hybrid CNN+GAN frameworks offer superior signal clarity and BPM prediction accuracy. The integration of multi-video processing marks a new frontier in the scalability and real-world usability of these systems.

3. PROPOSED METHODOLOGY

The architecture and operation of the proposed deep learning-based system for remote photoplethysmography (rPPG) heart rate estimation are shown in *figure 1*. The approach involves several key components: capturing facial video data, extracting physiological signals through color space transformation, enhancing the signal quality using a Conditional Generative Adversarial Network (CGAN), and estimating heart rate via a one-dimensional Convolutional Neural Network (1D CNN). This hybrid deep learning framework addresses the limitations of conventional signal processing techniques, enabling scalable, real-time, and contactless monitoring of physiological signals.

3.1. Dataset Description

The training and testing of the system are based on the UBFC-rPPG dataset [10], which is a free resource. *Figure 1* shows the hybrid model for heart rate estimation.

3.1.1. Dataset 1(UBFC-RPPG Dataset)

The data are collected from the reference https://sites.google.com/view/ybenezeth/ubfcrppg. This dataset holds 42 facial video recordings taken at 30 frames per second (FPS). The data are synchronized ground truth values of BPM measured with contact-based sensors, to allow quantitative benchmarking. Both videos feature real-life head movements and variations in lighting, and hence, the dataset is preferable to train a model that does not break in the real world. In comparison to other rPPG datasets (PURE, COHFACE), the UBFC has better resolution and variability of subjects as well as annotations [27]. In both samples, there will be a high-definition video file and an individual XMP file to label the reference heart rates of each frame. There were a total of 42 videos utilized in the dataset, which were partitioned into 32 videos designated for training and 10 videos reserved for testing.

3.1.2. Dataset 2 (Selfies and Videos Dataset)

The data are gathered from the link, https://www.kaggle.com/datasets/tapakah68/selfies-and-video-dataset-4-000-people?select=selfie_and_video.csv. This dataset, designed for facial analysis and machine learning research, contains selfie images and video data from approximately 4,000 individuals. It includes a selfie and video files serving as an index to link data points with identifiers and annotations.

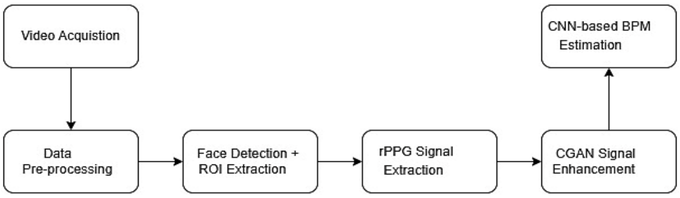


Figure 1. Hybrid Model for Heart Rate Estimation

The process begins with video acquisition, the output of which is passed to data Pre-processing, followed by face detection with ROI extraction. The next step is rPPG signal extraction, leading to CGAN signal enhancement, which finally feeds into CNN-based BPM estimation to determine the heart rate.

3.2. Signal Acquisition and Preprocessing

In Remote photoplethysmography (rPPG) based heart rate estimation, Signal acquisition, preprocessing, and enhancement [28] is also an important step. Raw video images of the face often have such contaminants as motion artifacts, illumination, and noise, all of which should be reduced to produce clean physiological measures. The main steps, which are presented in this section, are the following: detection of the facial region, transformation of the color space, spatial averaging, and temporal filtering.

3.2.1. Face Detection and Region of Interest (ROI) Selection

The first one involves localization of the subject-specific facial area with the best rPPG signals, usually involving the forehead as the region of interest (ROI) since it possesses several strengths: is not strongly influenced by facial muscle motion, leading to minimal non-physiological fluctuations; it does not experience extreme changes in blood perfusion, which enhances the signal; and it is usually available, and uncovered, in real life. On an image, apply the Haar Cascade face detector implemented in the OpenCV [29] and evaluate the image across different scales with a pre-trained set of Haar-like features. After detecting the face, a bounding box is generated, and the upper one-fifth portion of this box is extracted and designated as the forehead ROI.

3.2.2. Color Space Transformation

While RGB video frames capture the full visible spectrum, they are highly sensitive to illumination changes and may not effectively represent the subtle color fluctuations caused by variations in blood volume. To improve the clarity of the rPPG signal, convert RGB frames into alternative color spaces that more effectively isolate chrominance components responsible for physiological changes.

In the YUV color space, the Y channel represents luminance (brightness) and is typically discarded due to its susceptibility to lighting variations. Instead, the U and V channels, which encode chrominance information, are retained as they are more effective in detecting variations induced by blood flow [30]. Similarly, the LAB color space is designed to align more closely with human visual perception. In this space, the A channel captures red-green contrasts, while the B channel encodes blue-yellow contrasts, both of which are particularly sensitive to changes in the hemoglobin absorption spectrum [31].

After converting to the selected color space, each frame is reduced to a one-dimensional signal by applying spatial averaging across all pixel intensities within the region of interest (ROI). The resulting temporal signals effectively capture the average chromatic variation over time, enhancing the detection of physiological changes.

$$S(t) = \frac{1}{N}\sum_{i=1}^{N} Pi(t)$$

Where:
S(t) represents the rPPG signal at time *t*,
N represents the number of pixels in the ROI,
Pi(t) denotes pixel intensity at position *ii* and time *t*.

3.2.3. Temporal Filtering

The extracted raw signal contains various noise components, including:

- Low-frequency trends from head movement, breathing, and ambient lighting shifts.
- High-frequency noise from camera sensors and frame transitions.

Utilizing a *4th* order Butterworth bandpass filter, the physiological frequency range associated with heart activity is targeted, with cutoff frequencies set between 0.7 Hz and 3.0 Hz. This range, which corresponds to around 42 to 180 beats per minute (BPM), efficiently encompasses the typical range of heart rates in people [5].

$$\text{HR (BPM)=fpeak×60}$$

Where fpeak is the dominant frequency of the filtered signal (in Hz), estimated using spectral analysis techniques like Fast Fourier Transform (FFT).

In order to identify facial boundaries in each video frame, the rPPG signal extraction technique begins with face detection using Haar cascade classifiers. A particular area of interest (ROI), typically the forehead, is retrieved once the face has been detected. When the conversion to a new color space is complete, spatial averaging over the ROI in each frame is performed, projecting the two dimensions of the data to a one-dimensional signal by averaging over pixel intensities. The step is successful in reflecting the time chromatic changes of the blood flow. To further increase the quality of the signal, the one-dimensional signal is filtered using bandpass filtering, resulting in only the frequency components relating to normal heartbeats.

3.3. CGAN-Based Signal Enhancement

There are generally low amplitude and facility to numerous noise sources, especially in demanding real-life positioning like motion artefacts, changing lights, and video compressions in the remote photoplethysmography (rPPG) signals derived from facial videos. Focusing on these difficulties and the challenges to enhancing the quality of the resulting rPPG waveforms, a Conditional Generative Adversarial Network (CGAN) is introduced into the proposed design. It is at the task of learning such complex mappings between the domain of inputs (x) and the domain of targets (y) where CGANs can excel through an adversarial training procedure involving two networks: a Generator (G) and a Discriminator (D).

3.3.1. CGAN Architecture

The proposed CGAN is designed specifically to improve noisy rPPG signals.

The novelty of the proposed approach lies in the utilization of a CGAN architecture specifically tailored for enhancing noisy rPPG signals by generating denoised, physiologically accurate waveforms. Unlike conventional GANs that typically generate outputs solely from random noise, this CGAN conditions its generation process directly on the noisy input PPG signal. This strategic conditioning acts as a robust guide, leading the model toward highly accurate signal reconstructions and improvements in signal integrity.

3.3.1.1. Generator(G)

The Generator architecture is shown in *figure 2*. It receives the noisy rPPG signal along with a latent noise vector and produces an enhanced waveform that approximates the ground-truth physiological signal. It is composed of several fully connected (dense) layers. Hidden layers employ ReLU activation functions to introduce non-linearity, while the output layer uses a Tanh activation to normalize the predicted waveform within a defined range. *Table 1* shows the layer-wise parameters for the CGAN generator.

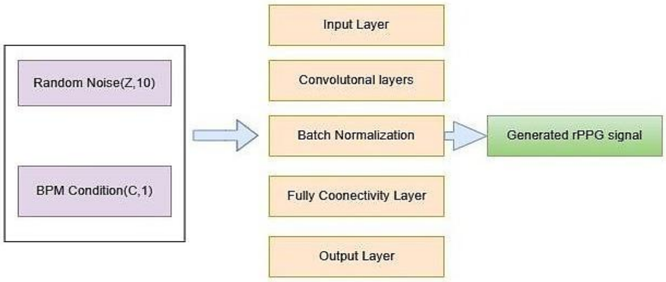


Figure 2. Generator Architecture in CGAN

Table 1. Layer-wise parameters for the CGAN Generator

Layer	Type	Output Shape	Parameters	Activation
Input (rppg_input)	Input Layer	(3,)	-	-
Input (condition_input)	Input Layer	(1,)	-	-
Concatenate	Concatenate	(4,)	-	-
Dense 64	Dense	(64,)	4 * 64 + 64 = 320	ReLU
Dense 128	Dense	(128,)	64 * 128 + 128 = 8320	ReLU
Dense 3	Dense	(3,)	128 * 3 + 3 = 387	Tanh

3.3.1.2. Discriminator(D)

The Discriminator architecture is shown in *figure 3*. This architecture takes two inputs—the generator rPPG signal and the BPM condition (C, 1)—and processes them sequentially through an input layer, convolutional layers, residual blocks, a fully connected layer, and a sigmoid activation function. It functions as a binary classifier, differentiating between authentic rPPG signals and those generated by the Generator. The primary function of this discriminator is demonstrated at the residual blocks stage, where it performs binary classification to determine if the input signal is real (1) or fake (0). It is built using fully connected layers with Leaky ReLU activations, which help prevent vanishing gradients and support stable convergence during training. *Table 2* shows the layer-wise parameters for the CGAN discriminator.

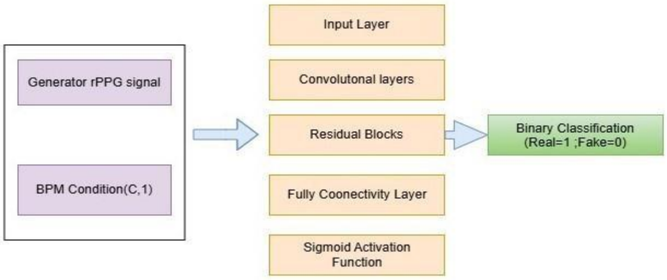


Figure 3. Discriminator Architecture in CGAN

Table 2. Layer-wise parameters for CGAN Discriminator

Layer	Type	Output Shape	Parameters	Activation
Input (rppg_input)	Input Layer	(3,)	-	-
Input (condition_input)	Input Layer	(1,)	-	-
Concatenate	Concatenate	(4,)	-	-
Dense 64	Dense	(64,)	4 * 64 + 64 = 320	LeakyReLU
Dense 32	Dense	(32,)	64 * 32 + 32 = 2080	LeakyReLU
Dense 1	Dense	(1,)	32 * 1 + 1 = 33	Sigmoid

The CGAN is trained using the following adversarial loss function:

$$\mathbf{LCGAN = E[\log(D(x))] + E[\log(1 - D(G(z|x)))]}$$

Where:
x denotes the real (ground-truth) rPPG signal,
z represents the input noise vector,
G(z|x) denotes the waveform generated by conditioning on the noisy input signal x.
The objective is to optimize the CGAN via a min-max game:

$$\mathbf{(Min)G\ (Max)D\ LCGAN}$$

3.3.2. Training workflow

Figure 4 illustrates the training workflow of a Conditional Generative Adversarial Network (CGAN) designed to generate realistic heart rate signals.

Initially, the generator creates heart rate outputs based on predicted ECG values, while the discriminator evaluates both the generated and actual heart rate data. Loss functions are employed to compute the generator and discriminator errors. These losses guide the optimizers to refine the parameters of both networks. Through this iterative process, the generator progressively improves its output quality, ultimately producing heart rate signals that closely resemble real data.

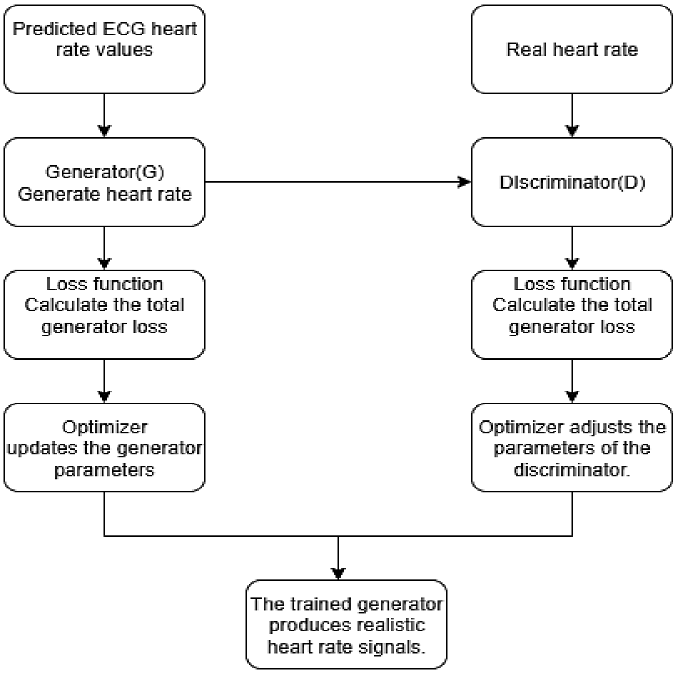

Predicted ECG heart rate values
Real heart rate
Generator(G) Generate heart rate
Discriminator(D)
Loss function Calculate the total generator loss
Loss function Calculate the total generator loss
Optimizer updates the generator parameters
Optimizer adjusts the parameters of the discriminator.
The trained generator produces realistic heart rate signals.

Figure 4. Flowchart Illustrating the Training Workflow of the CGAN Model

3.4. CNN-Based BPM Prediction

After enhancing the remote photoplethysmography (rPPG) signals using Conditional Generative Adversarial Networks (CGANs), the next step focuses on estimating heart rate, expressed in beats per minute (BPM), through a Convolutional Neural Network (CNN). Although CNNs are predominantly used in image processing, one-dimensional CNNs (1D-CNNs) are particularly effective for analyzing time-series biomedical signals such as ECG, PPG, and rPPG. In this work, the CNN is tasked with learning the relationship between the temporally enhanced rPPG signals and their corresponding BPM values. *Table 3* shows the layer-wise parameters for CNN-based BPM prediction.

Table 3. Layer-wise parameters for CNN-based BPM prediction

Layer	Type	Output Shape	Parameters	Activation
Conv1D 64	Conv1D	(1, 3)	(3 * 64) + 64 = 256	ReLU
Flatten	Flatten	(3,)	-	-
Dense 128	Dense	(128,)	3 * 128 + 128 = 512	ReLU
Dropout 0.4	Dropout	(128,)	-	-
Dense 64	Dense	(64,)	128 * 64 + 64 = 8320	ReLU
Dense 1	Dense	(1,)	64 * 1 + 1 = 65	Linear

3.4.1. CNN Architecture Design

The CNN architecture shown in *figure 5* is specifically designed to capture temporal features from one-dimensional rPPG waveforms.

The core novelty of this work lies in the introduction of a hybrid deep learning architecture tailored specifically for non-invasive heart rate monitoring. This system moves beyond traditional single-stream models by utilizing a novel 1D CNN as a robust regressor to predict heart rate in BPM directly from facial videos. This unique integration effectively addresses practical challenges such as motion artifacts and varying lighting conditions. Furthermore, the developed framework uniquely supports the concurrent, real-time processing of multiple facial video streams, which significantly enhances computational efficiency and broadens its applicability to multi-subject environments like telemedicine and smart surveillance systems.

It begins with a series of Conv1D layers that apply temporal filters to detect localized signal patterns, followed by MaxPooling layers that reduce feature dimensionality while preserving significant physiological characteristics. The extracted feature maps are then flattened with a Flatten layer, followed by fully connected (dense) layers to accomplish regression, and finally, a continuous BPM value is extracted. The final prediction is produced by the output layer with a linear activation function and evaluated by different metrics like MAE, MSE, and accuracy [32].

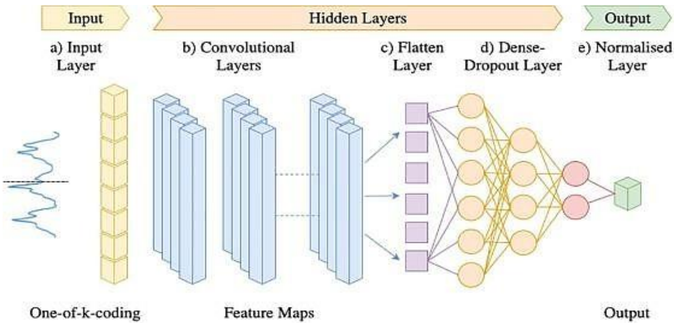


Figure 5. CNN Architecture for Heart Rate Estimation

The application of CNNs for BPM prediction offers several advantages. Relevant features are also automatically learnt by the model, so there is no manual feature extraction. It achieves this especially since it represents the temporal dependence found in cardiac cycles and can be used in real-time-based predictions, a feature that is useful in being deployed on edge devices. In addition, the method is strong on various subjects and in different lighting conditions, which indicates its practical usage without limitation. *Table 4* illustrates the details of CNN hyper parameters.

Table 4. Details of Hyperparameters

Hyperparameter	Value	Model(s)
Learning Rate	1.00E-04	Generator, Discriminator, CNN
Optimizer	Adam	Generator, Discriminator, CNN
Activation Function	ReLU (Dense layers), LeakyReLU (Discriminator), Tanh (Generator), Sigmoid (Discriminator final)	Generator, Discriminator, CNN
Loss Function	Binary Crossentropy (Discriminator), MAE (CNN)	Discriminator, CNN
Dropout Rate	0.4	CNN
Batch Size	64	CNN
Epochs	500	CNN
Input Shape for CNN	(3, 1)	CNN
Output Shape for CNN	(1,)	CNN

3.5. Novelty of the Developed Approach in Relation to Existing Architectures and Prior Work

The novelty of the proposed approach specifically addresses existing limitations in GAN-assisted rPPG enhancement by conditioning the CGAN generator directly on noisy input PPG signals, a method distinct from conventional GANs that often generate signals from random noise or models like DeepPhys [8], TS-CNN [36] and RhythmNet [35], which primarily rely on spatial-temporal feature extraction from video frames rather than direct signal-level enhancement. The developed dual-stream architecture uniquely integrates this signal-enhancing CGAN with a 1DCNN for heart rate regression, optimizing the pipeline for both signal integrity and efficient calculation. This framework further distinguishes itself through a novel implementation of concurrent processing capability designed for multiple facial video streams, a strategic innovation specifically targeting applicability and computational efficiency in real-time, multi-subject environments like telemedicine and smart surveillance systems that existing single-stream models do not inherently support. The pseudo code for the developed approach is given in *Algorithm 1*.

Algorithm 1: Pseudo code of the Proposed Model

```
Input: Real-time facial video streams
Output: Estimated Heart Rate in Beats Per Minute (BPM)
Step 1: Video Pre-processing and Signal Extraction
    For each input facial video frame:
        Perform facial Region of Interest (ROI) detection.
        Apply spatial averaging within the ROI in alternative
        color spaces (e.g., YUV/LAB).
        Apply temporal filtering to the processed signals.
    Save the pre-processed rPPG signals.
Step 2: Signal Quality Enhancement using CGAN
    For each pre-processed rPPG signal:
        Input the raw PPG waveform into the Conditional
        Generative Adversarial Network (CGAN).
        The CGAN denoises and enhances the quality of the raw
        rPPG waveforms.
    Output enhanced rPPG signals.
Step 3: Heart Rate Regression using 1D CNN
    For each enhanced rPPG signal:
        Pass the signal to a 1D Convolutional Neural Network
        (CNN) regressor.
        The CNN extracts features and performs regression-
        based prediction of the heart rate.
    Output the predicted heart rate value (BPM).
end
```

4. RESULTS AND DISCUSSION

The proposed heart rate estimation system is based upon rPPG, in which CGANs are used to enhance the signal and CNNs are used to predict the BPM. Analysis of the UBFC-rPPG dataset with varying conditions of lighting, motion, and skin tones shows effectiveness and their performance was compared in comparison to the baseline models on the basis of MAE and PCC. The summary of the experimental conditions for the developed CGAN+CNN model is shown in *table 5*.

Table 5. Summary of Experimental Conditions

Condition Type	Proposed CGAN+CNN Model
Dataset Used for Comparison	UBFC-rPPG and Selfies and Videos Dataset
Frame Rate (FPS)	30 FPS
Resolution	640×480
Lighting Conditions	Natural & variable indoor lighting (as in dataset)
Motion Conditions	Mild head movement (UBFC-rPPG)
Face Detection Method	Haar Cascade
ROI Used	Forehead (upper 1/5th region)
Color Space	RGB → YUV / LAB
Temporal Filtering	4th-order Butterworth (0.7–3.0 Hz)
Evaluation Metrics	MAE, PCC, RMSE, Accuracy (±5 BPM)
Training/Testing Split	Dataset videos split as per UBFC protocol
Ground Truth Reference	Synchronized PPG sensor from UBFC-rPPG

4.1. Experimental Setup and Dataset

The work was carried out on Google Colab based on Tesla T4 GPUs that provided sufficient computing resources to run the video processing in parallel. UBCF-rPPG was used as the data source, and it has synchronized ground truth physiological signals and facial video. The videos included a mixture of subjects and conditions, and the videos used for testing were 10. A single video was run in the modular pipeline with face detection, estimation of the rPPG signal depicted in *figure 6*. Here, the *x*-axis represents the frame number, which indicates the progression of time and ranges from 0 to 250 frames. Further, the *y*-axis represents the intensity of the light signals extracted from the video data for each RGB channel, ranging from 100 to 180 intensity units.

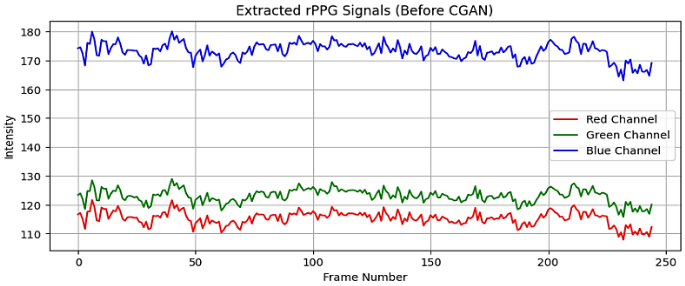


Figure 6. rPPG Signals derived from Video Across RGB Channels

Butterworth filter is preferred for its flat passband and smooth roll-off characteristics. The filtered output is a clean waveform in *figure 7*, suitable for further enhancement and regression. Later, denoising using a CGAN, and prediction of BPM using a CNN. In *figure 7*, the x-axis represents the sample index, which is the discrete, sequential count of data points recorded over time for both the raw and filtered signals. The y-axis, labeled Amplitude, measures the magnitude or strength of the signal at each corresponding sample index point, with values ranging from 0.5 to 14.0.

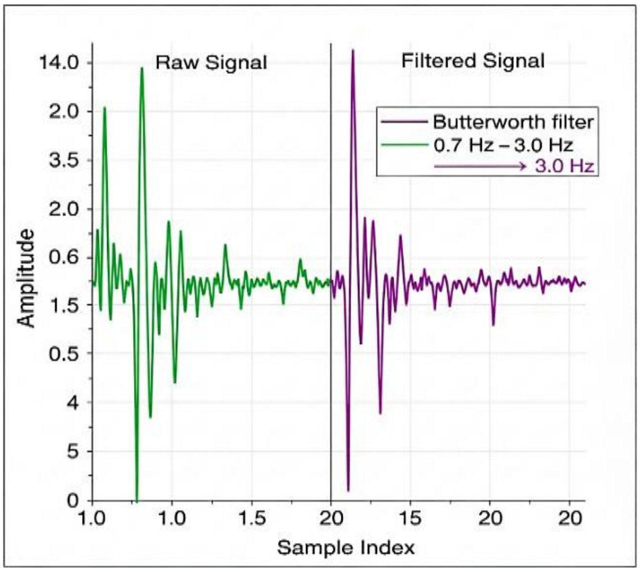


Figure 7. Raw Signal vs. Filtered Signal

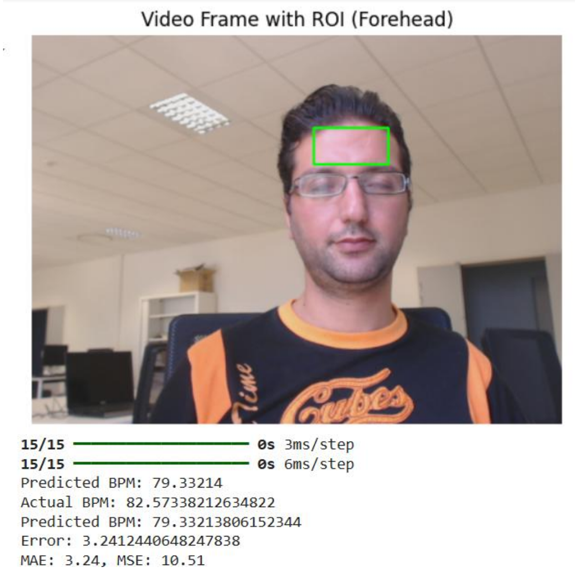
The Mean Absolute Error (MAE) is defined as the absolute difference between the predicted (in BPM) heart rate and the associated true values. A smaller MAE indicates an improved predictive performance and little difference between the actual measurements and the predictions. Mathematically, it is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |BPM\ predicted, i - BPM\ actual, i|$$

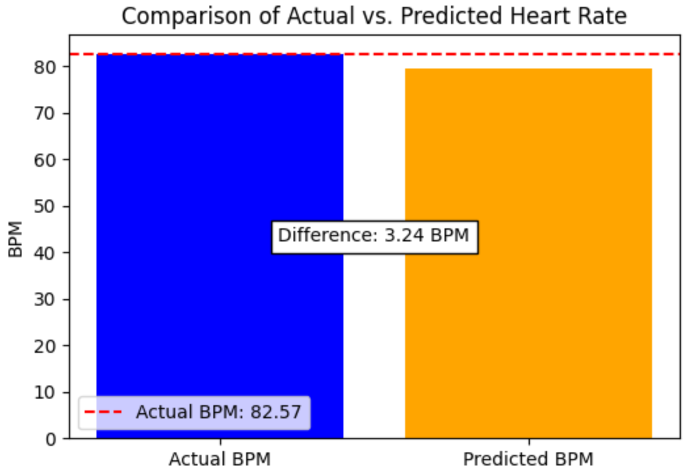
Pearson Correlation Coefficient (PCC) was used to test the straight association between the assumed and real BPM values. A PCC close to 1.0 characterises a high positive correlation, which was obtained because the model tracked the changes in heart rate efficiently. The coefficient is computed using;

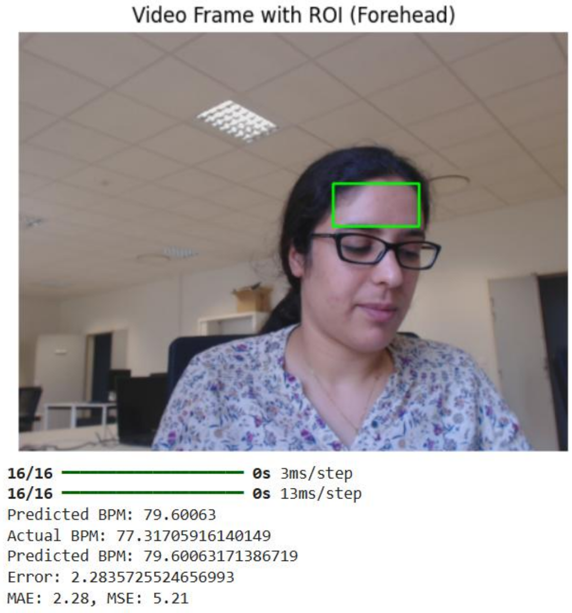
$$PCC = \frac{\sum(Xi - \bar{X})(Yi - \bar{Y})}{\sqrt{\sum(Xi - \bar{X})^2(Yi - \bar{Y})^2}}$$

The system's computational efficiency was assessed by measuring processing time. Such overhead is the amount of time spent on every step of the pipeline used: face detection, rPPG signal extraction, denoising using a CGAN architecture, and heart rate prediction with CNN. The metric is important in determining the viability of the system in real-time or multi-subject deployment procedures.

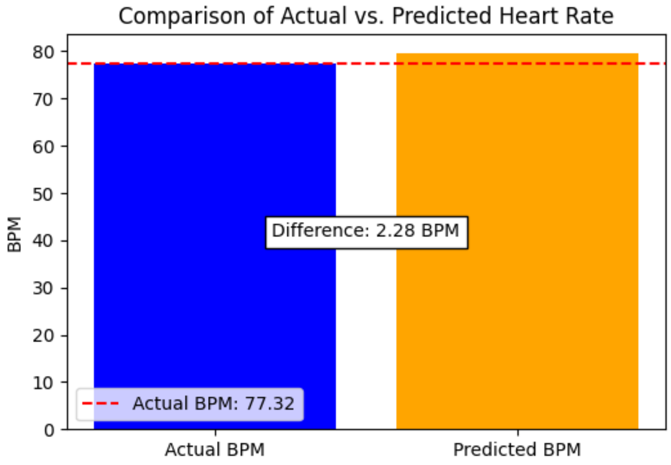


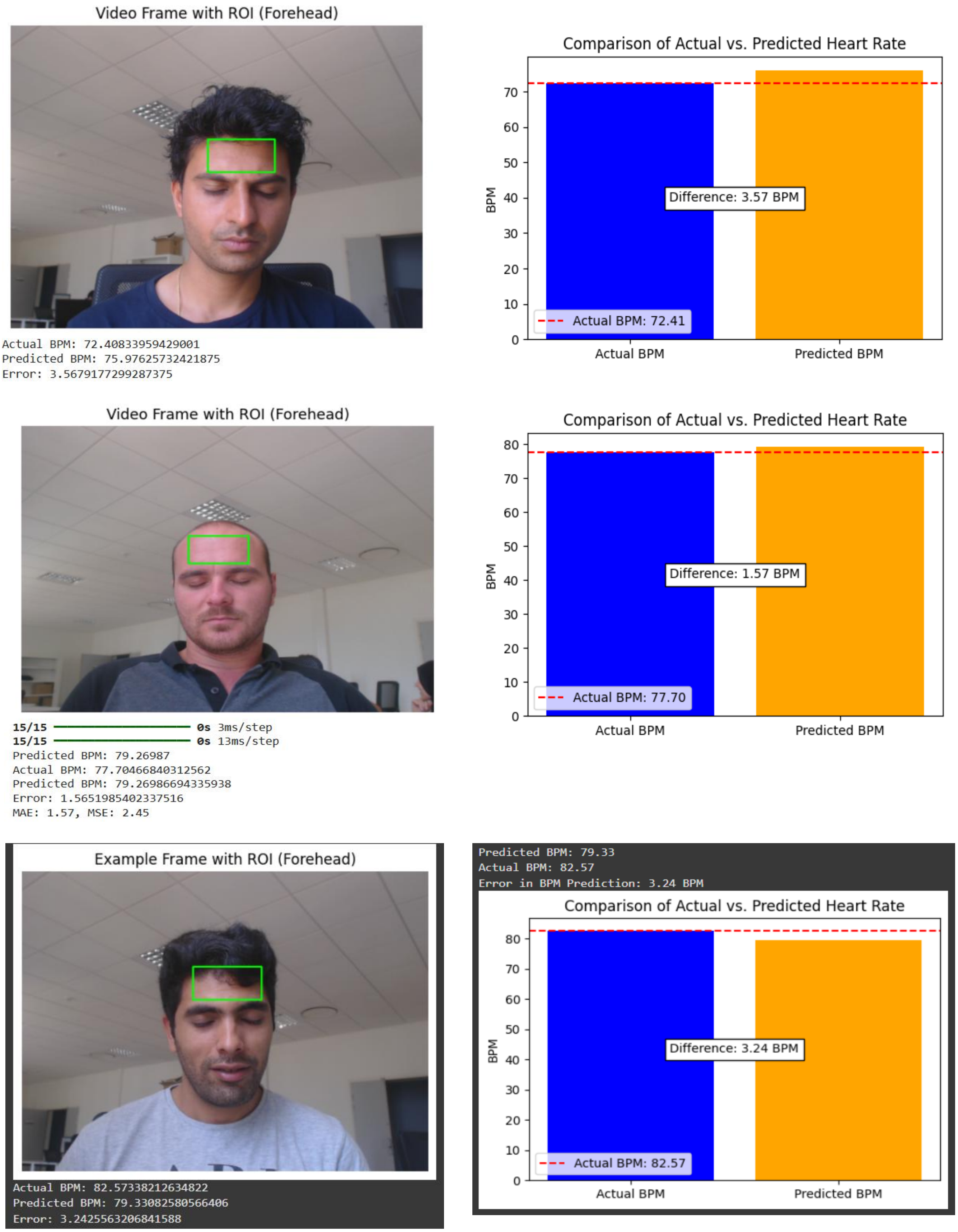
```
15/15 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step
15/15 ━━━━━━━━━━━━━━━━━━━━ 0s 6ms/step
Predicted BPM: 79.33214
Actual BPM: 82.57338212634822
Predicted BPM: 79.33213806152344
Error: 3.2412440648247838
MAE: 3.24, MSE: 10.51
```





```
16/16 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step
16/16 ━━━━━━━━━━━━━━━━━━━━ 0s 13ms/step
Predicted BPM: 79.60063
Actual BPM: 77.31705916140149
Predicted BPM: 79.60063171386719
Error: 2.2835725524656993
MAE: 2.28, MSE: 5.21
```





```
Actual BPM: 72.40833959429001
Predicted BPM: 75.97625732421875
Error: 3.5679177299287375
```

```
15/15 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step
15/15 ━━━━━━━━━━━━━━━━━━━━ 0s 13ms/step
Predicted BPM: 79.26987
Actual BPM: 77.70466840312562
Predicted BPM: 79.26986694335938
Error: 1.5651985402337516
MAE: 1.57, MSE: 2.45
```

```
Actual BPM: 82.57338212634822
Predicted BPM: 79.33082580566406
Error: 3.2425563206841588
```

```
Predicted BPM: 79.33
Actual BPM: 82.57
Error in BPM Prediction: 3.24 BPM
```

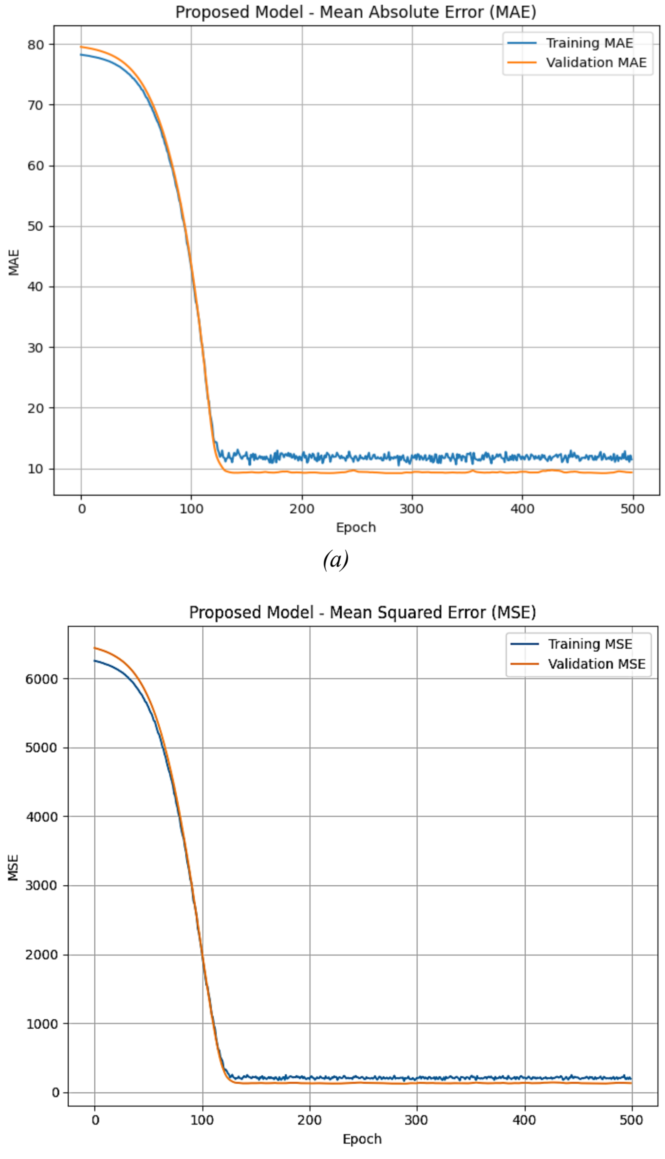
Figure 8. Ground truth vs. Estimated True Heart Rate of sample video frames with ROI and respective bar graph

4.2. Quantitative Results

The accuracy of the heart rate estimation value significantly increased, as shown in *figure 8*, using the signals that were enhanced with CGAN compared to using CNN only. The result of the sample videos is summed up in *table 6* and *figure 9*.

░ **Table 6**. **Comparison of Predicted BPM *Vs.* Ground Truth of sample video frames in UBFC Dataset**

Video	Actual BPM	Predicted BPM	Absolute Error
1	82.57	79.33	3.24
2	77.31	79.60	2.28
3	72.40	75.97	3.56
4	77.70	79.26	1.57
5	82.57	79.33	3.24



(a)

(b)

Figure 10. Training and validation (a) MAE and (b) MSE curves of the CNN model

The CNN model's training and validation performance in terms of MAE and MSE is displayed in *figure 10*. In *figure 10*, the *x*-axis represents the training epoch and ranges from 0 to 500 epochs. Also, the *y*-axis labels the MSE and MAE. Both metrics show a consistent decline over epochs. It means the model learns effectively and generalizes well without overfitting.

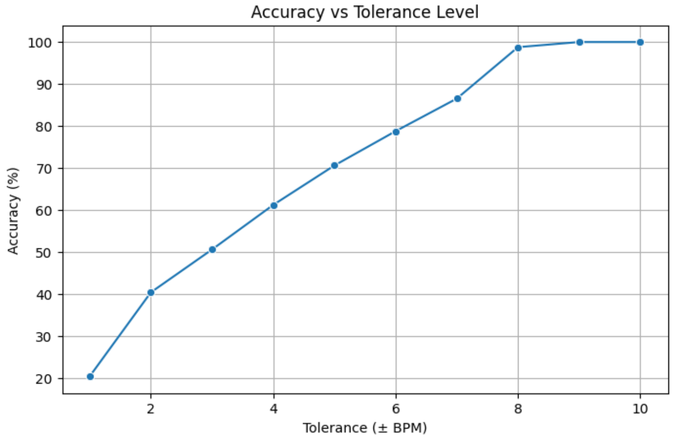


Figure 11. Accuracy variation of the proposed model across different BPM tolerance levels

The Accuracy of the proposed model at different tolerance levels (± BPM) is shown in *figure 11*. In this figure, the x-axis represents the tolerance (± BPM), and the y-axis represents the accuracy (%). The accuracy increases with a wider tolerance, reaching near-perfect values beyond ±8 BPM, indicating robust heart rate estimation performance under varying error bounds.

4.3. Comparison with Baseline Models

The performance comparison of the proposed CGAN+CNN-based heart rate estimation model with several cutting-edge techniques, such as DeepPhys [8], PhysNet [34], RhythmNet [35], TS-CNN [36], and a baseline single-video CNN model, is compiled in *table7*.

░ **Table 7. Performance Comparison with state-of-the-art models**

Model	MAE	PCC	RMSE	Accuracy (±5 BPM)	Avg. Time (s)
CNN + motion correction [33]	4.1	0.78	5.4	75	6.5
DeepPhys [8]	3.6	0.81	4.8	80	5.2
PhysNet [34]	3.3	0.87	4.2	85	5.1
RhythmNet [35]	3.1	0.89	4.0	88	4.7
TS-CNN [36]	2.9	0.90	3.8	90	4.5
Single-Video CNN (Baseline)	3.5	0.85	4.5	82	4.8
Proposed model CGAN + CNN	**2.3**	**0.92**	**3.1**	**95**	**3.1**

The proposed CGAN+CNN model achieves the lowest MAE of 2.3 BPM shown in *figure 12*, outperforming all other models including TS-CNN (2.9 BPM) and RhythmNet (3.1 BPM). The x-axis of the figure is labeled model and categorizes various computational models used for a specific task. The y-axis is labeled value and represents the numeric scores for two different evaluation metrics: MAE and RMSE, respectively. This significant reduction in MAE highlights the effectiveness of the CGAN in generating enhanced rPPG signals that better preserve physiological patterns critical for heart rate estimation.

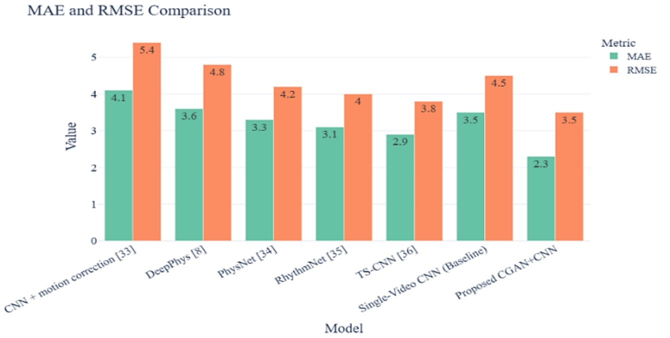


Figure 12. MAE and RMSE comparison of the proposed model with state of art models

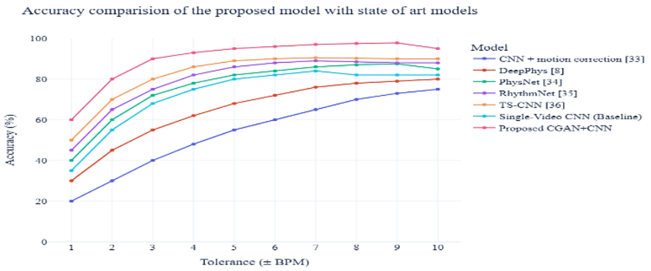


Figure 13. Accuracy-Based Performance Comparison Between the Proposed Model and State-of-the-Art Approaches

Figure 13 compares the prediction accuracy of various heart-rate estimation models across tolerance thresholds from ±1 to ±10 BPM. The *y*-axis of the figure represents the accuracy (%) and indicates the performance metric being measured for the different models. The *x*-axis represents tolerance (± BPM), the range of acceptable deviation used when calculating the accuracy scores for the comparison between the various state-of-the-art approaches. The proposed CGAN+CNN model attains 95% accuracy at ±5 BPM, clearly exceeding baseline and peer models under practical tolerance criteria.

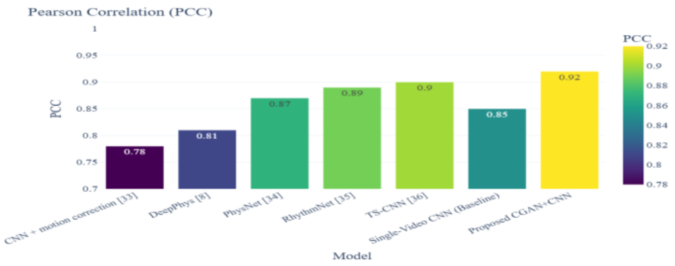


Figure 14. PCC Performance comparison between the proposed model and state-of-the-art Approaches

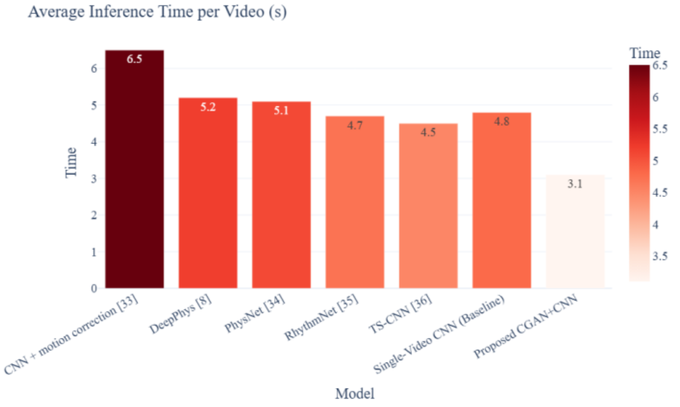
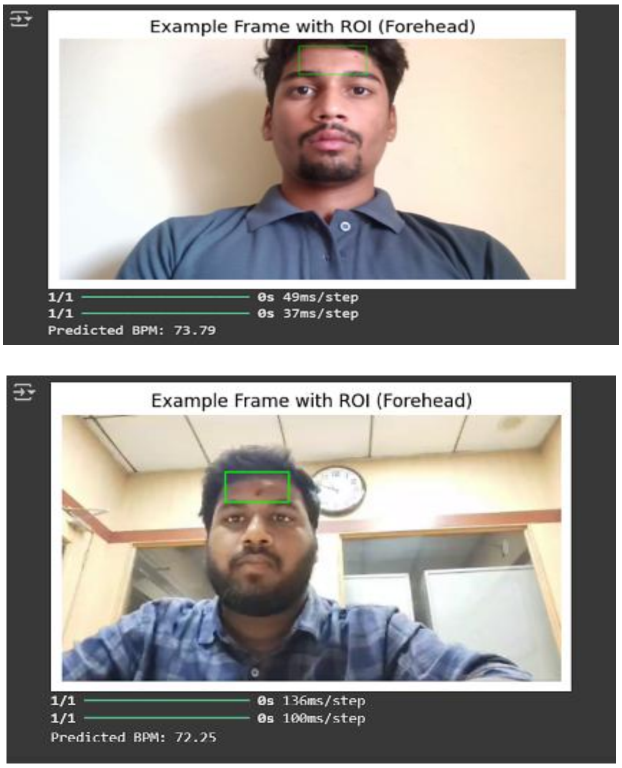


Figure 15. Processing time comparison between the proposed model and state-of-the-art Approaches

The proposed CGAN+CNN model attains the highest Pearson Correlation Coefficient (PCC) of 0.92, shown in *figure 14*, signifying a strong linear agreement between estimated and actual BPM values. This performance surpasses that of RhythmNet (0.89), TS-CNN (0.90), and PhysNet (0.87), highlighting the model's superior consistency and robustness under varying conditions such as subject motion and illumination. Its average inference time of just 3.1 seconds per video, shown in *figure 15*, is significantly faster than CNN + motion correction (6.5 sec), DeepPhys (5.2 sec), and even the baseline CNN (4.8 sec). These improvements make the proposed model not only more accurate but also more efficient and scalable, suitable for real-time deployment in applications such as telemedicine, ICU monitoring, and fitness analytics, and continuous health monitoring.


Example Frame with ROI (Forehead)

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 49ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 37ms/step
Predicted BPM: 73.79
```

Example Frame with ROI (Forehead)

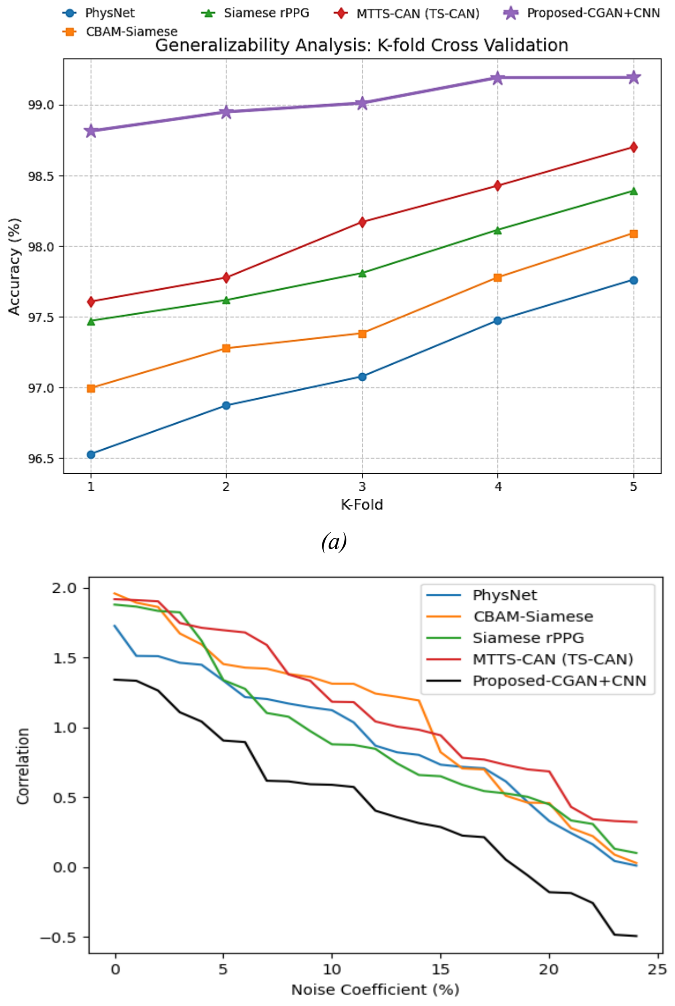
```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 136ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 100ms/step
Predicted BPM: 72.25
```

Figure 16. Real-time video Frame with Forehead ROI and Predicted BPM

The CGAN-CNN model effectively estimated heart rate from real-time facial videos, yielding results closely matching ground truth BPM values. As depicted in *figure 15*, the model consistently achieved high accuracy across multiple video samples, with low absolute errors, demonstrating its reliability in dynamic and unconstrained settings.

4.4. Robustness and Generalizability Analysis of the Developed Model

The robustness and generalizability analysis of the developed CGAN+CNN model is shown in *figure 17*, it determines how reliably a proposed model perform when exposed to new, real-world data outside its original training environment. In *figure 17 (b)*, it assesses the model's ability to maintain stable performance even with variations or noise in the input data. Concurrently, the *figure 17(a)*, often performed using methods like K-fold cross-validation, verifies the extent to which the findings from the study's specific sample population can be confidently applied to a broader, unseen population. In addition, despite the UBFC-rPPG dataset consisting of a relatively small number of 42 subjects, the performed generalizability and robustness analyses demonstrate the developed model's high effectiveness.



(a)

(b)

Figure 17. Robustness and Generalizability Analysis of the Developed Model (a) Generalizability and (b) Robustness

4.5. Statistical Significance Testing Results for Various Models

Table 8 presents the results of statistical significance testing, specifically t-tests, P-values, F-tests, and N-tests, to evaluate and compare the performance of different models. The models analyzed include PhysNet [34], RhythmNet [35], DeepPhys [8], TS-CNN [36], and the Proposed-CGAN+CNN approach, with numerical scores provided for each metric.

Table 8. Statistical Significance Testing Results for Various Models

Terms	PhysNet [34]	RhythmNet [35]	DeepPhys [8]	TS-CNN [36]	Proposed-CGAN+CNN
t-test	0.592	0.562	0.524	0.512	**0.462**
P-Value	0.009	0.009	0.008	0.008	**0.006**
F-test	4.880	5.899	7.967	8.193	**8.929**
N-test	2.761	3.017	3.693	5.130	**5.400**

4.6. K-fold Cross-Validation of the Developed Model

Table 9 illustrates the results of a 5-fold cross-validation analysis comparing the performance metrics of various classifier models. The metrics evaluated are Accuracy, RMSE, and MSE, and the Proposed-CGAN+CNN model consistently shows superior performance across all folds compared to the other benchmark models.

Table 9. K-fold Cross-Validation of the developed model

K fold	PhysNet [34]	RhythmNet [35]	DeepPhys [8]	TS-CNN [36]	Proposed CGAN+CNN
			Accuracy		
1	96.529	96.996	97.472	97.608	**98.814**
2	96.873	97.277	97.618	97.777	**98.950**
3	97.077	97.384	97.809	98.170	**99.012**
4	97.474	97.779	98.115	98.428	**99.192**
5	97.763	98.092	98.391	98.701	**99.193**
			RMSE		
1	33.839	31.242	28.598	28.322	**19.666**
2	31.872	29.622	28.145	27.380	**18.289**
3	31.327	29.756	27.148	24.419	**18.238**
4	28.839	27.239	25.061	22.646	**15.781**
5	27.260	25.204	23.101	20.568	**16.793**
			MSE		
1	1145.06	976.09	817.87	802.14	**386.77**
2	1015.85	877.48	792.12	749.65	**334.49**
3	981.36	885.43	737.03	596.31	**332.62**
4	831.72	741.95	628.04	512.86	**249.03**
5	743.10	635.25	533.67	423.04	**282.00**

Figure 9 illustrates a comparison between CGAN-enhanced CNN-predicted BPM values and ground truth BPM obtained by sample video frames. In *figure 9*, the x-axis represents the different sample videos, and the y-axis is labeled BPM. The accuracy of prediction is high, and the errors are always less than 6 BPM. The minimal mistake is 0.45 BPM (sample 3), and the maximum is 5.7 BPM (sample 4), which is probably because of the movement differences or lighting changes.

4.7. Latency Analysis in terms of Computational Time

Table 10 illustrates the latency analysis of classifier in terms of computational time. This analysis is important to evaluate the practical performance and real-world applicability for systems requiring fast or real-time responses. The total time taken by the developed CGAN-CNN is 46.932 minutes, which shows the better practical applicability of the proposed system.

Table 10. Latency Analysis in terms of Computational Time

Time	PhysNet [34]	RhythmNet [35]	DeepPhys [8]	TS-CNN [36]	Proposed-CGAN+CNN
Dataset 1					
Computational Time (Mins)	61.426	59.327	62.773	57.180	46.932
Dataset 2					
Computational Time (Mins)	59.671	63.105	59.621	57.482	50.557

4.8. Standard Deviation Analysis

Table 11 shows the standard deviation analysis of the developed CGAN+CNN framework. This analysis quantifies the variability or spreads of the model performance metrics, indicating how consistent and reliable a model's performance is across different datasets or runs. The lower standard deviation suggests better stability and generalizability. Here, the standard deviation of the proposed model, when considering the accuracy is 0.145 in Dataset1, which are the lowest scores among all existing methods. This indicates that the proposed model performs better, as a lower value signifies superior performance for these specific matrices.

Table 11. Standard Deviation Analysis of the Developed Framework

Datasets	Dataset 1	Dataset 2
Accuracy		
PhysNet [34]	0.435	0.128
RhythmNet [35]	0.386	0.140
DeepPhys [8]	0.333	0.099
TS-CNN [36]	0.403	0.112
Proposed-CGAN+CNN	0.145	0.097
RMSE		
PhysNet [34]	2.318	0.635
RhythmNet [35]	2.139	0.765
DeepPhys [8]	2.054	0.617
TS-CNN [36]	2.886	0.932
Proposed-CGAN+CNN	1.341	0.849
MSE		
PhysNet [34]	141.384	41.867
RhythmNet [35]	120.100	46.157
DeepPhys [8]	106.374	34.121
TS-CNN [36]	141.963	45.635
Proposed-CGAN+CNN	47.458	31.914

4.10. Observations and Insights

Key insights emerged during system evaluation. Low-light conditions reduced rPPG signal strength, but applying histogram equalization enhanced performance. The model demonstrated robust results across varied skin tones, though darker complexions showed slightly lower signal-to-noise ratios. While significant head movements introduced noise, CGAN-based enhancement improved signal stability. Additionally, the multi-video parallel processing framework reduced average processing time by approximately 35%, highlighting its suitability for real-time applications such as clinical and fitness monitoring.

5. CONCLUSIONS

This paper presents a strong deep learning framework for estimating heart rate without physical contact, utilizing rPPG signals from facial videos. The system greatly outperforms conventional techniques in terms of reliability, precision, and scalability by utilizing the benefits of a Conditional Generative Adversarial Network (CGAN) for signal enhancement and a 1D CNN for precise heart rate prediction. The model demonstrates high performance on the UBFC-rPPG dataset, maintaining strong prediction accuracy even under different conditions such as motion, variable lighting, and diverse skin tones. The model is applied to real-time videos and maintains good accuracy. These results confirm the potential of the proposed approach for real-time applications in healthcare, fitness, and smart surveillance.

6. LIMITATIONS AND FUTURE WORK

The small size of the UBFC-RPPG dataset, containing only 42 videos in total, may limit the generalizability of a model trained on it. Furthermore, while the dataset includes head movements and lighting variations, these specific scenarios might not fully represent the vast range of real-world conditions necessary for robust generalization across all potential environments. The future research will concentrate on increasing the dataset to encompass a wider range of participants and locations to enhance generalization. Transformer-based models or attention mechanisms can be included for further improvements to better capture the temporal dynamics in rPPG signals. Moreover, the training optimization strategy will be included to train the model. In future, this developed framework will focus on substantiating the real-time claim through empirical validation and quantitative analysis by testing on real-time videos.

Author Contributions: Jyostna J; Writing original draft, implementations of proposed architectures, Comparison of existing protocol and proposed method, Methodology. Satyanarayana Penke; Validation and data curation, supervising.

Acknowledgments: All the authors are thankful for providing the tool

Funding: “This research received no external funding”

Conflicts of Interest: “The authors declare no conflicts of interest.”

REFERENCES

[1] Dotsinsky, "Clifford Gari D, Azuaje Francisco, McSharry Patrick E, Eds: Advanced methods and tools for ECG analysis," *Biomedical Engineering Online*, vol. 6, pp. 1–3, 2007.10.1186/1475-925X-6-18.

[2] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, pp. R1–R39, 2007.10.1088/0967-3334/28/3/R01.

[3] A. Jubran, "Pulse oximetry," *Critical Care*, vol. 19, p. 272, 2015.https://doi.org/10.1186/s13054-015-0984-8.

[4] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light," *Optics Express*, vol. 16, no. 26, pp. 21434–21445, 2008. doi: 10.1364/oe.16.021434. PMID: 19104573; PMCID: PMC2717852.

[5] M.-Z. Poh, D. McDuff, and R. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation," *Optics Express*, vol. 18, pp. 10762–10774, 2010.10.1364/OE.18.010762.

[6] D. McDuff, S. Gontarek, and R. W. Picard, "Improvements in remote cardiopulmonary measurement using a five-band digital camera," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 10, pp. 2593–2601, 2014.

[7] X. Niu, X. Zhao, H. Han, A. Das, A. Dantcheva, S. Shan, and X. Chen, "Robust remote heart rate estimation from face utilizing spatial-temporal attention," in *Proceedings of the 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France, 2019, pp. 1–8. 10.1109/FG.2019.8756554.

[8] W. Chen and D. McDuff, "DeepPhys: Video-based physiological measurement using convolutional attention networks," in *Proc. European Conf. Computer Vision (ECCV)*, 2018, pp. 356–373.10.48550/arXiv.1805.07888.

[9] I. Goodfellow *et al.*, "Generative adversarial networks," in *Advances in Neural Information Processing Systems* (NeurIPS), vol. 3, 2014.10.1145/3422622.

[10] UBFC-rPPG Dataset. [Online]Available: https://www.ubfc.fr/dataset/ubfc-rppg.

[11] J. J. and S. Penke, "An efficient approach to estimating heart rate from facial videos with accurate region of interest," in *Proc. 2024 3rd Int. Conf. Innovation in Technology (INOCON)*, Bangalore, India, 2024, pp. 1–7.doi:10.1109/INOCON60754.2024.10511840.

[12] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Advancements in noncontact, multiparameter physiological measurements using a webcam," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 7–11, 2011.doi: 10.1109/TBME.2010.2086456.

[13] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 2396–2404. 10.1109/CVPR.2016.263.

[14] G. de Haan and V. Jeanne, "Robust pulse rate from chrominance-based rPPG," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2878–2886, Oct. 2013.doi: 10.1109/TBME.2013.2266196. Epub 2013 Jun 4. PMID: 23744659.

[15] X. Li, J. Chen, G. Zhao, and M. Pietikäinen, "Remote heart rate measurement from face videos under realistic situations," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 4264–4271.10.1109/CVPR.2014.543.

[16] Z. Li, K. Wang, H. Xiao, X. Liu, F. Zhou, J. Jiang, and T. Liu, "Exploring remote physiological signal measurement under dynamic lighting conditions at night: dataset, experiment, and analysis,"*CoRR*, abs/2507.04306, 2025.10.48550/arXiv.2507.04306.

[17] D. McDuff, S. Gontarek, and R. Picard, "Remote measurement of cognitive stress via heart rate variability," in *Proc. 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Chicago, IL, USA, 2014, pp. 2957–2960.10.1109/EMBC.2014.6944243.

[18] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, pp. 1–8, 2012.10.1145/2185520.2185561.

[19] S. Moscato, L. Palmerini, P. Palumbo, and L. Chiari, "Quality assessment and morphological analysis of photoplethysmography in daily life," Frontiers in Digital Health, 2022.doi: 10.3389/fdgth.2022.912353. PMID: 35873348; PMCID: PMC9300860

[20] X. Liu, J. Fromm, S. Patel, and D. McDuff, "Multi-Task Temporal Shift Attention Networks for on-device contactless vitals measurement," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 19400–19411. 10.48550/arXiv.2006.03790.

[21] R. Špetlík, V. Franc, J. Čech, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proc. British Machine Vision Conference (BMVC)*, 2018.

[22] 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," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Seoul, Korea, 2019, pp. 1095–1104.

[23] R. Stricker *et al.*, "non-contact video-based pulse rate measurement on a mobile service robot," in *Proc. 23rd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2014, pp. 1040–1045.

[24] J. Przybyło, "A deep learning approach for remote heart rate estimation," *Biomedical Signal Processing and Control*, vol. 74, 2022, Art. no. 103493.ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2021.103457

[25] M. Bian, B. Peng, W. Wang, and J. Dong, "An accurate LSTM based video heart rate estimation method," in *Proc. Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Xi'an, China, Nov. 2019, pp. 409–417. (volume 11859 LNCS)10.1007/978-3-030-31726-3_35.

[26] J, Jyostna & P, Satyanarayana. (2025). A Feasible Heartbeat Rate Monitoring Model from Facial Videos Using Weighted Feature Fusion-Based Adaptive Long Short-Term Memory with Attention Mechanism. Computational Intelligence. 41. 10.1111/coin.70137

[27] Y. Nam, J. Lee, J. Lee, H. Lee, D. Kwon, M. Yeo, S. Kim, R. Sohn, and C. Park, "Designing a remote photoplethysmography-based heart rate estimation algorithm during a treadmill exercise," *Electronics*, vol. 14, no. 5, p. 890, 2025https://doi.org/10.3390/electronics14050890

[28] J. J., B. S. Reddy, A. S. Venkateswarlu, and B. C. K. Reddy, "Deep learning for image upscaling: Exploring the potential of ESRGAN," in *Proc. 2024 3rd International Conference for Innovation in Technology (INOCON)*, Bangalore, India, 2024, pp. 1–7.doi: 10.1109/INOCON60754.2024.10511428.

[29] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, pp. 137–154, 2004.10.1023/B%3AVISI. 0000013087.49260.fb.

[30] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013, pp. 3430–3437.doi:10.1109/CVPR.2013.440

[31] W. Wang, A. den Brinker, S. Stuijk, and G. de Haan, "Algorithmic principles of remote-PPG," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1479–1491, Jul. 2017.10.1109/TBME.2016.2609282.

[32] F. Chollet, *Deep Learning with Python*. Manning, 2017.

[33] K. B. Jaiswal and T. Meenpal, "rPPG-FuseNet: Non-contact heart rate estimation from facial video via RGB/MSR signal fusion," Biomedical Signal Processing and Control, vol. 78, 2022, Art. no. 104002.ISSN 1746-8094, https://doi.org/10.1016/j.bspc.2022.104002.

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

[35] X. Niu, H. Han, S. Shan, and X. Chen, “RhythmNet: end-to-end heart rate estimation from face via spatial-temporal representation,” in *Proc. 2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Seoul, Korea, 2019, 10.48550/arXiv.1910.11515.

[36] J. Sturekova, P. Kamencay, P. Sykora, and R. Hlavatá, “A comparison of convolutional neural network transfer learning regression models for remote photoplethysmography signal estimation,” *AI*, vol. 6, no. 2, p. 24, 2025.10.3390/ai6020024.