Journal of Information Systems Engineering and Management
2025, 10(10s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Developing Multi-Channel Pulse Signal Analysis for Enhanced Heart Rate Detection Using Facial Video Systems

Jyoti Vitthal Chhatrband[1], Dr. Bhavesh Kumar Choithram Dharmani[2]
[1,2]School of electronics and electrical engineering, Lovely Professional University, Punjab, India
Email: jojitagurav@gmail.com[1], Email: dharmanibc@gmail.com[2]

ARTICLE INFO

Received: 09 Nov 2024

Revised: 27 Dec 2024

Accepted: 13 Jan 2025

ABSTRACT

By leveraging advanced signal processing techniques and illumination correction, the study enhances the accuracy and robustness of heart rate detection systems under varying lighting conditions. This paper introduces the development of the Unified Pulse Detection from Complex Environments (UPDCE) Model, a deep learning framework designed for the non-invasive detection of heart rate from facial video data. Utilizing the UBFC-RPPG dataset, which includes video recordings under various illumination conditions, the model employs convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to analyze multiple specific regions of interest namely, the forehead and chin. These areas are critical for capturing pulse signals influenced by subtle changes in skin coloration due to blood flow. The model processes video frames, extracted at a three-second interval, through stages of enhancement and normalization to improve data quality for subsequent analysis. Features are then extracted and temporally analyzed to detect and calculate heart rate accurately. Special emphasis is placed on overcoming challenges associated with diverse lighting and motion conditions. The system architecture ensures robust processing by incorporating techniques to optimize real-time operation and reduce computational load. The effectiveness of the UPDCE Model is validated through rigorous training and testing, demonstrating significant potential for real-world application in continuous health monitoring systems. This research contributes to advancements in remote photoplethysmography by highlighting methodological innovations and deployment strategies that enhance the accuracy and reliability of heart rate detection using facial video analysis.

Keywords: Remote Photoplethysmography, Heart Rate Detection, Deep Learning, Facial Video Analysis, Signal Processing, Health Monitoring Systems, Illumination Correction, Real-Time Video Processing

I. Introduction

Heart rate monitoring is a critical component of modern health diagnostics, providing essential data for evaluating cardiac health, physical fitness, and overall physiological stress levels. Traditionally, heart rate measurement has relied on direct contact methods, such as electrocardiography (ECG) and photoplethysmography (PPG), which, while accurate, can be inconvenient and restrictive for continuous monitoring. As technology has progressed, there has been a critical thrust towards creating non-invasive methods that can offer comparable exactness without the require for coordinate physical contact [1]. This paper presents a novel approach to non-invasive heart rate checking by creating and pre-processing multi-channel beat signals extricated from confront video investigation, centering particularly on the regions of the chin and forehead. The utilization of confront recordings for heart rate location is predicated on the optical assimilation characteristics of human skin. When lit up, light enters the skin and is retained by the blood vessels, which change in volume with each pulse. These unobtrusive changes in light assimilation can be captured through video and hence analyzed to extricate the beat signal—a handle known as inaccessible photoplethysmography (rPPG). In any case, the viability of this procedure is intensely subordinate on different components counting lighting conditions, the subject's development, and the physiological contrasts in skin and tissue composition over people [2]. These variables can present noteworthy commotion and changeability into the captured signals, in this manner challenging the unwavering quality of the heart rate estimations derived from face recordings. This paper presents an advanced multi-channel approach that upgrades the strength and precision of heart rate location from confront recordings [3]. By analyzing different districts of intrigued inside the video particularly the forehead and chin areas our show totals differing beat signals that shift in their affectability to movement and lighting artifacts. This multi-channel approach not as it were increments the sum of data available for examination but moreover gives an implies to cross-verify the extricated heart rates, in this manner progressing the certainty within the estimations [4]. The flag quality is assist refined through an arrangement of pre-processing steps planned to adjust and normalize light varieties over the video outlines. Light adjustment is vital, as conflicting lighting can skew the intensity values seen within the video, driving to mistakes within the beat flag extraction. Our method employs adaptive histogram equalization and other picture handling procedures to stabilize lighting conditions over the video sequence [5], guaranteeing that the beat signals are inferred from physiological changes instead of environmental artifacts. The advancement of the multi-channel beat flag show includes several stages, beginning with the extraction of raw data from high-resolution video captures [6]. Using advanced facial recognition algorithms, the system identifies and isolates the regions of interest on the subject's face—namely, the forehead and chin. These areas are chosen due to their relative immobility compared to other facial features, and their rich capillary compositions, which enhance the detectability of pulse signals.

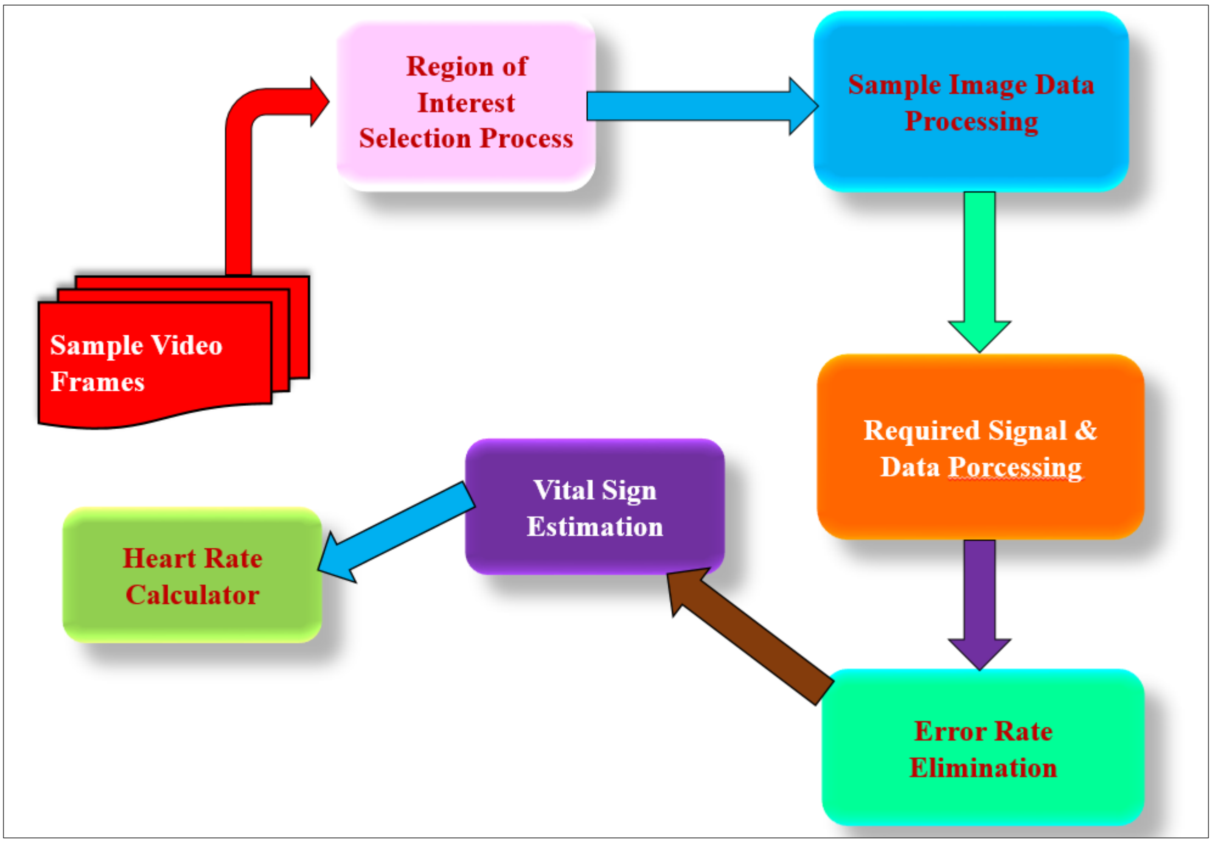

Region of Interest Selection Process
Sample Image Data Processing
Sample Video Frames
Required Signal & Data Porcessing
Vital Sign Estimation
Heart Rate Calculator
Error Rate Elimination

Figure 1. Illustrate the Basic Overview Blcok Diagram of Heart Rate System Analyser

Following the isolation of these areas, the model applies a series of digital filters to each channel to mitigate the impact of random noise and non-pulse-related variations, such as those caused by the subject's minor movements or changes in facial expression [7]. The filtered signals from each channel are at that point synchronized and combined employing a weighted calculation that optimizes the commitment of each channel based on its flag clarity and soundness. The integration of signals from numerous channels requires the utilize of modern information combination methods. This investigate leverages machine learning calculations, especially convolutional neural systems (CNNs) and recurrent neural networks (RNNs), to memorize from the worldly and spatial designs of the beat signals [8]. These systems are prepared on a assorted dataset comprising confront recordings recorded beneath different natural conditions to upgrade the model's flexibility and precision in real-world scenarios as shown in Figure 1. This multi-channel, machine learning-enhanced system speaks to a critical progression within the field of non-invasive heart rate checking. It not as it were addresses the restrictions of current rPPG strategies but too sets a modern standard for precision and unwavering quality in inaccessible wellbeing observing innovations. Through comprehensive testing and approval [9], this consider points to illustrate the common sense of utilizing confront video examination for heart rate discovery, clearing the way for its integration into clinical hones and buyer wellbeing applications. This paper is organized to begin with detail the technique utilized in creating the multi-channel beat flag show, taken after by a thorough assessment of its execution against ordinary heart rate monitoring methods. The subsequent sections will discuss the implications of these findings for future research and potential applications in both healthcare and fitness industries.

II. Literature Review

Advancements in biomedical image segmentation and physiological measurement using computer vision have progressed significantly. Early methods, like U-Net, set a foundation for accurately segmenting complex biomedical images, leading to improved feature localization in medical imaging and physiological monitoring applications [10]. The integration of attention components, as seen in Squeeze-and-Excitation Systems, empowered these models to prioritize pertinent highlights, advance refining heart rate estimation frameworks. Lightweight models like GhostNet empowered real-time preparing, crucial for applications such as farther heart rate observing. Building on this, systems like EVM-CNN illustrated the achievability of contactless heart rate estimation from facial recordings by leveraging photo plethysmography (PPG) signals, highlighting the common sense of non-invasive physiological observing. Inquire about has too centered on the impact of video encoding on heart rate precision, underscoring the have to be keep up video quality for solid PPG [11]. With clean ground truth information and assessment systems, real-time webcam-based heart rate estimation has built up benchmarks for surveying framework reliability over different conditions. Recent approaches, such as contrastive learning, empowered heart rate estimation from unlabelled video information by recognizing unobtrusive contrasts in facial blood stream. This advance was advance upgraded with Contrast-Phys, which leverages spatiotemporal data for made strides flag strength against commotion. The presentation of consideration components through the Transformer show has been transformative, permitting frameworks to specifically center on basic transient and spatial highlights, in this manner progressing physiological flag extraction. Additionally, unsupervised skin segmentation techniques improved rPPG accuracy by isolating skin regions, enhancing the precision of contactless measurements [12]. Non-video-based methods, such as continuous-wave Doppler radar and microwave sensors, offer viable alternatives for specific applications, including driver monitoring. Transformer-based models like Phys Former have since emerged, capturing temporal dependencies in video-based heart rate measurements, and setting new standards for extracting physiological signals [13]. A detailed understanding of heart rate variability metrics remains foundational for monitoring stress-related changes, with applications in fitness and diagnostics. End-to-end solutions like AutoHR leverage neural architecture search to optimize heart rate measurement, making them adaptable for real-world conditions. Applications of machine learning in physiological monitoring also extend to specialized assessments [14], like autism diagnosis, where social visual attention analysis offers insights into behavioral health.

Research Focus	**Research Approach**	**Major Insights**	**Key Obstacles**	**Limitations**	**Potential Uses**
Biomedical Image Segmentation	U-Net: Convolutional Network	High accuracy in segmenting biomedical images	Requires large labeled datasets	High computational cost	Biomedical imaging
Feature Enhancement in CNN	Squeeze-and-Excitation Networks	Enhanced feature representation in networks	Added complexity to network design	Increased model size	Heart rate estimation
Lightweight CNN Architecture	GhostNet	Efficient feature extraction with low-cost operations	Limited to specific architectures	May sacrifice accuracy for speed	Real-time heart rate monitoring

Contactless Heart Rate Estimation	EVM-CNN	Effective for real-time contactless heart rate estimation	Sensitive to lighting conditions	Lower accuracy in low light	Remote heart rate monitoring
Video-based Physiological Measurement	Analysis of video encoding effects	Video compression artifacts affect physiological signal accuracy	Maintaining video quality	Limited to high-quality video	Camera-based heart rate estimation
Webcam Heart Rate Measurement	Webcam-based PPG with clean ground truth data	Improved accuracy with controlled data	Variability in lighting conditions	Limited scalability to uncontrolled environments	Real-time heart rate variability measurement
Contrastive Learning for rPPG	Unsupervised contrastive learning	Improved model learning in unsupervised settings	Requires complex contrastive data preparation	Requires large computational resources	Remote photoplethysmography
Unsupervised rPPG Measurement	Contrast-Phys using spatiotemporal contrast	Enhanced robustness in video-based physiological measurement	Sensitivity to large facial movements	Limited by motion artifacts	Unsupervised heart rate measurement
Attention Mechanisms in DL	Transformer model	Attention improves feature focus in temporal and spatial domains	High computational cost	High training cost	Diverse deep learning tasks

Table 1. Summarizes the Literature Review of Various Authors

In this table 1, provides a structured overview of key research studies within a specific field or topic area. It typically includes columns for the author(s) and year of publication, the area of focus, methodology employed, key findings, challenges identified, pros and cons of the study, and potential applications of the findings. Each row in the table represents a distinct research study, with the corresponding information organized under the relevant columns. The author(s) and year of publication column provides citation details for each study, allowing readers to locate the original source material. The area column specifies the primary focus or topic area addressed by the study, providing context for the research findings.

III. Model Design Methodology

The proposed system, as block diagram shown in Figure 2, for detecting heart rate using facial video data. We integrates several components, such as dataset storage, cloud storage, a web server, and a deep learning model employing CNN and RNN for feature extraction and temporal analysis, respectively. The system processes video data to calculate and display heart rates, identifying specific regions like the chin and forehead.

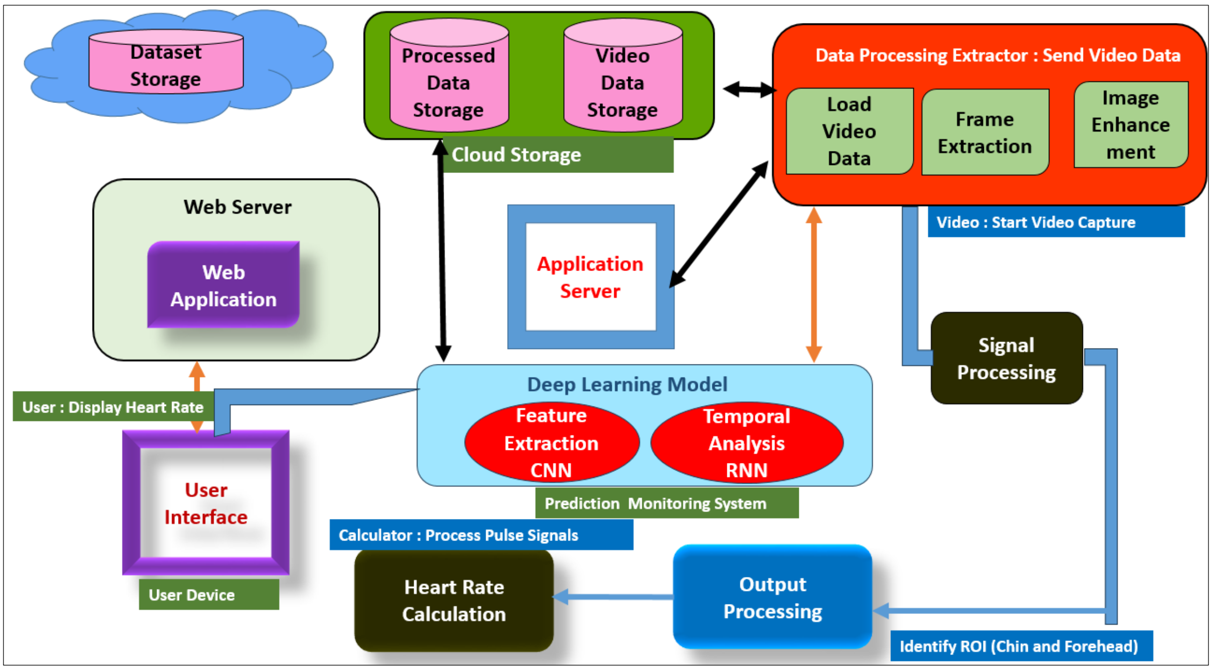


Figure 2. Block Design of Proposed Methodology

Stage -1] Input UBFC-rPPG Dataset:

In the first phase of proposed methodology involves the comprehensive collection of face video data, which serves as the foundational dataset for our analysis, from the dataset UBFC-rPPG [22]. This dataset comprises high-definition video recordings of subjects over a different statistic, counting varieties in age, skin tone, and sexual orientation. Recordings are captured in a controlled environment with flexible lighting conditions to simulate distinctive real-world scenarios, the sample images of dataset shown in Figure 3. The video and ECG information are time-stamped to encourage exact synchronization amid the flag preparing stage.

Stage -2] Region of Interest (ROI) Identification

With the dataset input, the step is to distinguish particular districts of intrigued (ROIs) on the subjects' faces, essentially centering on the brow and chin ranges. These districts are chosen based on their vascular properties, which are more conducive to recognizing unobtrusive changes in blood stream, and their relative movement soundness amid recordings. Progressed facial acknowledgment calculations are utilized to distinguish and separate these ROIs naturally in each video outline.

Stage -3] Signal Extraction

Once ROIs are characterized, the another arrange includes extricating the crude pulse signals from these ranges. This can be fulfilled through a advanced flag handling method known as inaccessible photo plethysmography (rPPG). The rPPG strategy analyzes the light assimilation and reflection characteristics of the skin as captured within the video. By centering on the green channel of the RGB video information, where the differentiate due to blood stream is most articulated, we improve the affectability of our beat discovery. The flag extraction prepare applies a band-pass channel to the crude information to evacuate high-frequency clamor (such as unpretentious lighting glints) and low-frequency floats (such as moderate changes in surrounding lighting). The sifted flag from each outline is at that point totaled to create a ceaseless flag waveform for each ROI over the term of the video.

Stage -4] Illumination Correction

To address varieties in lighting, which can altogether affect the precision of rPPG signals, we actualize an light rectification calculation. This calculation alters the intensity of the video outlines to a standardized lighting demonstrate, lessening the impact of shadows, glares, and other lighting artifacts. Methods such as versatile histogram equalization are utilized to make strides the consistency of lighting over all video outlines. This normalization guarantees that varieties within the extracted pulse signals are due to physiological changes or maybe than natural components.

Stage -5] Multi-Channel Signal Processing

The center of our strategy is the multi-channel investigation, where signals from multiple ROIs are coordinates to upgrade the exactness of the recognized heart rate. Each ROI produces an free beat flag, which can be influenced in an unexpected way by different clamor variables such as unpretentious developments or changes in facial expressions. To synthesize a single, robust heart rate measurement from multiple signals, we employ a weighted fusion algorithm. This algorithm assigns a reliability score to each channel based on the signal-to-noise ratio and the temporal consistency of the detected pulse. Signals with higher reliability scores are given more weight in the fusion process. The final heart rate is calculated using a peak detection algorithm applied to the fused signal, identifying the periodic peaks corresponding to each heartbeat.

Stage -6] Machine Learning Enhancement

To advance refine our heart rate location, machine learning models are prepared to distinguish and rectify peculiarities within the beat signals. A convolutional neural network (CNN) is utilized to memorize highlight representations from sections of the rPPG signals, whereas a repetitive neural arrange (RNN) with LSTM units models the transient elements of these highlights. The systems are prepared on a subset of our dataset that has been physically looked into and clarified for flag inconsistencies and heart rate mistakes. The machine learning models work in two stages: feature learning and decision-making. Within the include learning arrange, the CNN extricates spatial highlights from the time-series information of each signal, which are at that point sequenced through the RNN to capture transient designs. Within the decision-making organize, the arrange yields an adjustment figure for the heart rate based on recognized irregularities, progressing the by and large precision and unwavering quality of the framework.

Stage -7] Validation

At last, the execution of our created framework is approved against the ground truth ECG information. Measurable measures such as root cruel square blunder (RMSE), cruel outright blunder (MAE), and Pearson relationship coefficient are calculated to assess the exactness and consistency of the heart rate estimations inferred from our show. This comprehensive strategy guarantees that our approach to creating and pre-processing multi-channel beat signals for heart rate detection utilizing confront video investigation is strong, adaptable, and competent of delivering high-accuracy comes about over different real-world conditions.

IV. Existing Models for Video Analysis Detection

Video analysis, especially for detection tasks, employs a variety of models and techniques that are tailored to the specific needs of the application, whether it be object detection, activity recognition, or anomaly detection. Here are several prominent models and approaches commonly used in the field:

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are specialized sorts of neural systems that exceed expectations in handling information that includes a grid-like topology, such as pictures. A CNN regularly comprises of a grouping of layers that change the input volume into an yield volume through differentiable capacities. These layers incorporate convolutional layers, enactment capacities like ReLU, pooling layers, and completely associated layers that come full circle in an yield layer regularly organized for classification. The center building square of a CNN is the convolutional layer that applies a set of learnable channels to the input. Each channel in a convolutional layer is little spatially (along width and tallness), but amplifies through the total profundity of the input volume. For illustration, a normal channel on a to begin with layer of a CNN handling color pictures might have estimate 5x5x3 (5 pixels width, 5 pixels tallness, and 3 color channels). As the filter slides over the image, it produces a two-dimensional actuation outline that gives the reactions of that channel at each spatial position, capturing spatial pecking orders in information. Naturally, the arrange learns channels that activate when they see a few particular sort of highlight at a few spatial position within the input. Taking after the convolutional layers, pooling (subsampling or down examining) layers decrease the measurements of the information by combining the yields of neuron clusters at one layer into a single neuron within the following layer. Common pooling methods incorporate max pooling, which takes the most extreme esteem of a specific highlight over the fix of the image prepared by the channel, viably making the input representations littler and more reasonable.

CNNs are comprised of one or more layers that specifically include convolutional layers, pooling layers, and fully connected layers at the end:

1. **Convolutional Layers**: These layers perform a convolution operation that filters the input to extract features.

$$= (f * g)(i, j) = \sum m \sum n f(m, n) \cdot g(i - m, j - n)$$

2. **Activation Function**: Following the convolution operation, an activation function like ReLU (Rectified Linear Unit) is typically applied to introduce non-linearities into the model, helping it to learn more complex patterns.

$$h(x) = max(0, x)$$

3. **Pooling Layers**: After feature extraction, pooling (also known as subsampling or down sampling) reduces the dimensionality of each feature map but retains the most important information. Pooling layers summarize the features present in a region of the Average pooling takes the average of elements in a region of the feature map.

$$\boldsymbol{Max\ Pooling}: y = max\ (m, n) \in Rxm, ny$$

$$\boldsymbol{Average\ Pooling}: y = 1 \mid R \mid \sum(m, n) \in Rxm,$$

4. **Fully Connected Layers**: After several convolutional and pooling layers, the high-level reasoning in the neural network is done via fully connected layers. Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular neural networks. Their outputs can be calculated with a matrix multiplication followed by a bias offset.

$$y = Wx + b$$

Loss Function $\sigma(zi) = \sum j = 1Kezj/ezi$

5. **Input Layer**: The input layer takes the raw pixel values of the image.

$$L(y, p) = -\sum i = 1Cyilog(pi)$$

6. **Feature Learning**: Through multiple convolutional and max pooling layers, the network learns to identify various features of the image. Early layers may identify simple features like edges and textures, while deeper layers can identify high-level features like shapes or objects.

After passing through several convolutional, ReLU, and pooling layers, the high-level reasoning is done by fully connected layers. The last fully connected layer (often followed by a softmax activation function) provides the output probabilities for each class.

B. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are a specialized sort of neural arrange outlined to handle grouping expectation issues. Not at all like standard feedforward neural systems, RNNs are characterized by their capacity to preserve a frame of inside memory or state, which permits them to handle inputs of changing lengths and to capture transient flow successfully. This highlight makes RNNs especially well-suited for errands such as discourse acknowledgment, dialect modeling, and time arrangement expectation where the arrange and setting of information focuses are basic. RNNs accomplish this usefulness through their special engineering, where associations between hubs shape a coordinated chart along a transient arrangement.

Step 1: Initialize Network Parameters: Initialize hidden_state to zero Total loss = 0 Process input data xtx_txt which is part of a sequence x1,x2,...,xTx_1, x_2, ..., x_Tx1,x2,...,xT

Step 2: Sequence Input: input_t = X[t] true_output_t = Y[t]

Step 3: Forward Pass: hidden_state = activation_function(W_ih * input_t + W_hh * hidden_state + b_h) output_t = output_function(W_ho * hidden_state + b_o)

Step 4: Compute Loss: loss_t = cross_entropy_loss(output_t, true_output_t) Total_loss += loss_t **Step 5: Backward Pass (Backpropagation Through Time - BPTT)** :output_t = derivative of loss with respect to output_t W_ho, dL/db_o = backpropagate through output layer hidden_state, dL/dW_ih, dL/dW_hh, dL/db_h = backpropagate through time using hidden states

Step 6: Iterate: Repeat steps 3 through 5 for several epochs or until the model converges (i.e., the change in loss between epochs is minimal or below a threshold).

Step 7: Model Evaluation : Assess the performance of the model on a validation set using appropriate metrics (e.g., accuracy for classification, MSE for regression). Make adjustments to the model or training process based on the performance metrics.

Step 8: Deployment

W_ho -= learning_rate * dL/dW_ho b_o -= learning_rate * dL/db_o

W_ih -= learning_rate * dL/dW_ih

W_hh -= learning_rate * dL/dW_hh b_h -= learning_rate * dL/db_h

V. Proposed UPDCE Model (Deep Learning Model)

To create a deep learning model named "UPDCE Model" for the purpose of processing and analyzing face video for heart rate detection from multiple regions (such as the chin and forehead), we can outline a conceptual model based on convolutional neural networks (CNNs) and possibly recurrent neural networks (RNNs) if temporal analysis adds value. Here’s a step-by-step guide on how to design this model:

Step 1: Load Dataset

The UPDCE (Unified Pulse Detection from Complex Environments) Model aims to detect and analyze pulse signals from video data of the face, focusing on multiple specific regions to measure heart rate accurately even in diverse lighting and motion scenarios as displayed in Figure 3.

- **Dataset**: UBFC-RPPG dataset [22], which is a publicly available source used for remote photoplethysmography (rPPG) studies and includes face video recordings under different illumination conditions.

Figure 3. Dataset Video Samples

- **Operation**: Download the dataset from the provided UBFC-RPPG dataset link. Review the dataset documentation to understand the video specifications and annotations included.

Step 2: Extract Frames from Video

In this stage, we employed YOLO to select the facial region, and then extracted the forehead and chin areas from it, as these regions typically show the most uniform intensity distribution. These areas are likely to better represent color variations corresponding to changes in heart rate. The dataset should include regions of interest on the face (e.g., forehead and chin) and corresponding heart rate measurements obtained via a reliable method like an ECG as displayed in Figure 4.

- **Input**: Use the video files from the UBFC-RPPG dataset.

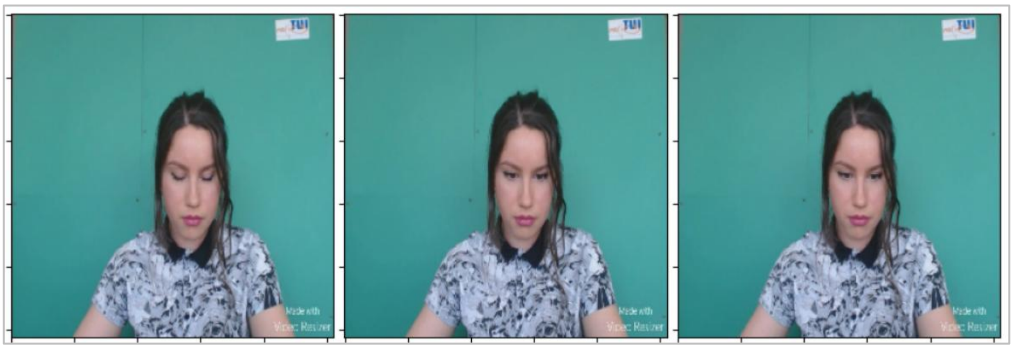


Figure 4. Extracted Frames from Video

- **Frame Generation Interval**: We have used the frame extraction to occur every 3 seconds of video time to manage the computational load and reduce redundancy in frames.
- **Output Folder**: We have created directory structure on the local machine or server to store the extracted frames.
- **Processing:**
    - We have load each video sequentially.
    - Next we have taken each video, and iterate it through the frames with a 3-second interval.
    - We did the resize of each selected frame to 256x256 pixels to standardize the input size for subsequent processing.
    - At last we save the resized frames to the designated output folder.

YOLO face Detection Algorithm Step wise

YOLO Model for Face Detection

Step 1: Model Architecture

- Grid Size: S×S

- Bounding Boxes per Cell: B

- Outputs per Box: (x, y, w, h, confidence)

- Output Dimensions: (S×S×B)×(5+C)

- where C = 1 for face detection.

Step 2: Bounding Box Prediction

- Normalized Coordinates (x, y):

$$x = \sigma(tx) + cx$$

$$y = \sigma(ty) + cy$$

- σ: sigmoid function
- tx, ty: model outputs
- cx, cy: grid cell top-left coordinates.

- Box Dimensions (w, h):

$$w = pw * exp(tw)$$

$$h = ph * exp(th)$$

- pw, ph: predefined anchor dimensions, tw, th: model outputs.

- Confidence Score:

$$Confidence = \sigma(to)$$

to: object presence prediction.

Step 3: Class Prediction

- Class Confidence (face):

$$Class\ Confidence = \sigma(tc) \times Confidence$$

- tc: class score output by the model.

Step 4: Loss Function

- Combined loss:

$$L = \lambda coord * \sum[(xi - x'i)^2 + (yi - y'i)^2] + \lambda obj * \sum(Ci - C'i)^\wedge 2 + \lambda class * \sum(pci - p'ci)^\wedge 2$$

- 1ijobj: indicator of object presence in box j of cell i.
- λ: constants to balance loss components.

Step 5: Non-max Suppression

- Filter excess bounding boxes using:

- Confidence threshold.

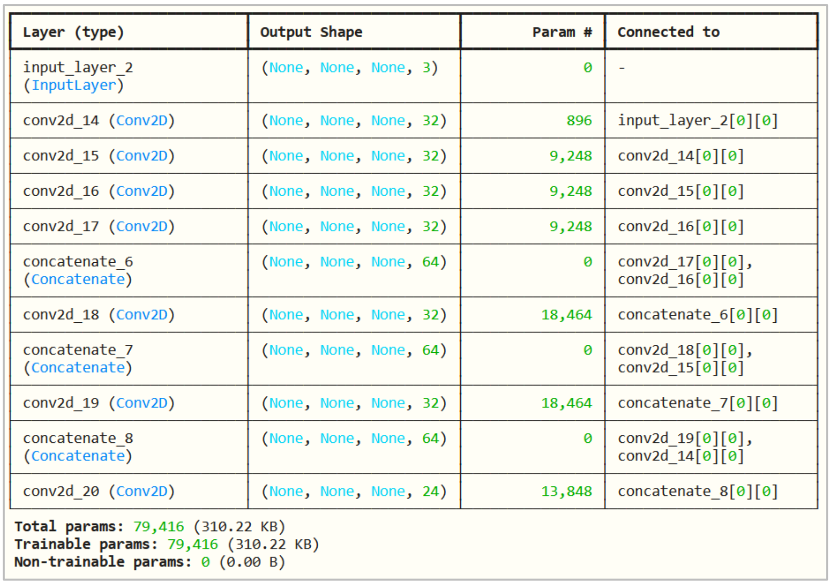
- IOU (Intersection Over Union) threshold; keep box with highest confidence if IOU > threshold.

$$- IOU\ Equation: IOU = \frac{Area\ of\ Overlap}{Area\ of\ Union}$$

Step 3: Frame/Image Enhancement

Automatically detected and extracted the forehead and chin areas using facial landmarks. We apply the image processing techniques to correct for lighting variations within these regions to standardize the input data for the neural network. Standardize the video frame intensities to have similar scales across the dataset as described in Table 2.

Table 2. Proposed UPDCE Model for Heart Rate Detection from Face Video



Layer (type)	Output Shape	Param #	Connected to
input_layer_2 (InputLayer)	(None, None, None, 3)	0	-
conv2d_14 (Conv2D)	(None, None, None, 32)	896	input_layer_2[0][0]
conv2d_15 (Conv2D)	(None, None, None, 32)	9,248	conv2d_14[0][0]
conv2d_16 (Conv2D)	(None, None, None, 32)	9,248	conv2d_15[0][0]
conv2d_17 (Conv2D)	(None, None, None, 32)	9,248	conv2d_16[0][0]
concatenate_6 (Concatenate)	(None, None, None, 64)	0	conv2d_17[0][0], conv2d_16[0][0]
conv2d_18 (Conv2D)	(None, None, None, 32)	18,464	concatenate_6[0][0]
concatenate_7 (Concatenate)	(None, None, None, 64)	0	conv2d_18[0][0], conv2d_15[0][0]
conv2d_19 (Conv2D)	(None, None, None, 32)	18,464	concatenate_7[0][0]
concatenate_8 (Concatenate)	(None, None, None, 64)	0	conv2d_19[0][0], conv2d_14[0][0]
conv2d_20 (Conv2D)	(None, None, None, 24)	13,848	concatenate_8[0][0]

```
Total params: 79,416 (310.22 KB)
Trainable params: 79,416 (310.22 KB)
Non-trainable params: 0 (0.00 B)
```

- **Model Creation**: Develop the UPDCE Model, a deep learning model tailored for enhancing images extracted from the videos.

- **Performance Parameters**: Configure the model to optimize based on:
  - Color constancy loss: Minimizes color shifts caused by lighting variations.
  - Exposure loss: Adjusts the exposure levels to avoid under or overexposure.
  - Illumination smoothness loss: Ensures even lighting within the frame.
  - Spatial Consistency Loss: Maintains spatial consistency across the enhanced frames.
- **Training and Testing**: Divide the dataset into training and testing portions. Use the training data to fit the model and validate its performance on the test set.
- **Image Enhancement**:
  - Process each frame through the trained UPDCE Model to enhance image quality as displayed in Figure 5.

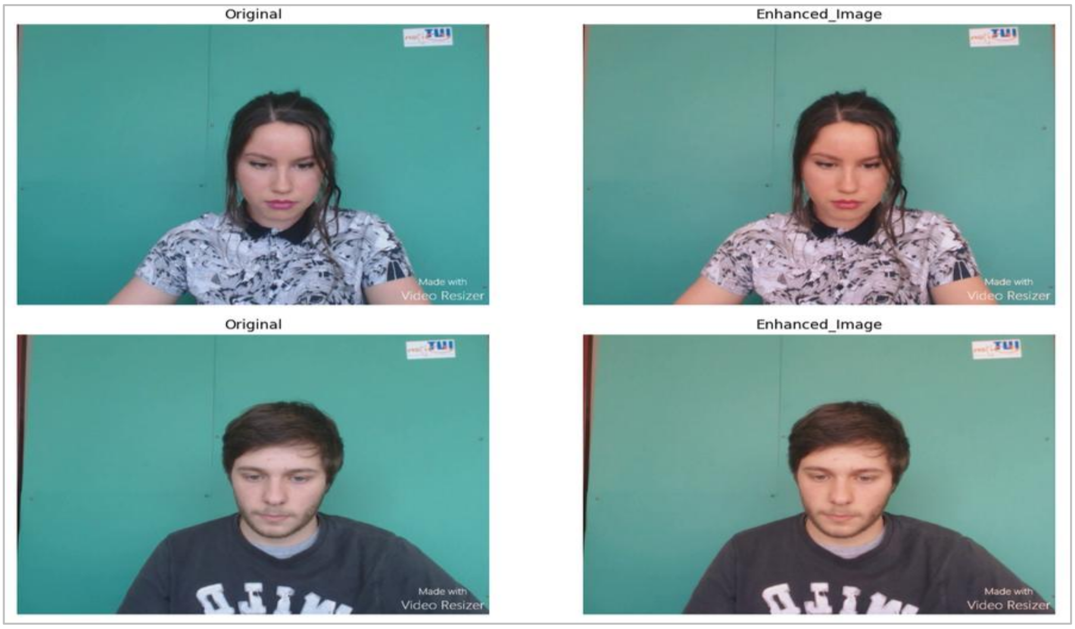


Figure 5. (a) Original Image (b) Enhance Image

  - Apply auto-contrast to the enhanced images using the Python Imaging Library (PIL) to improve visual clarity .

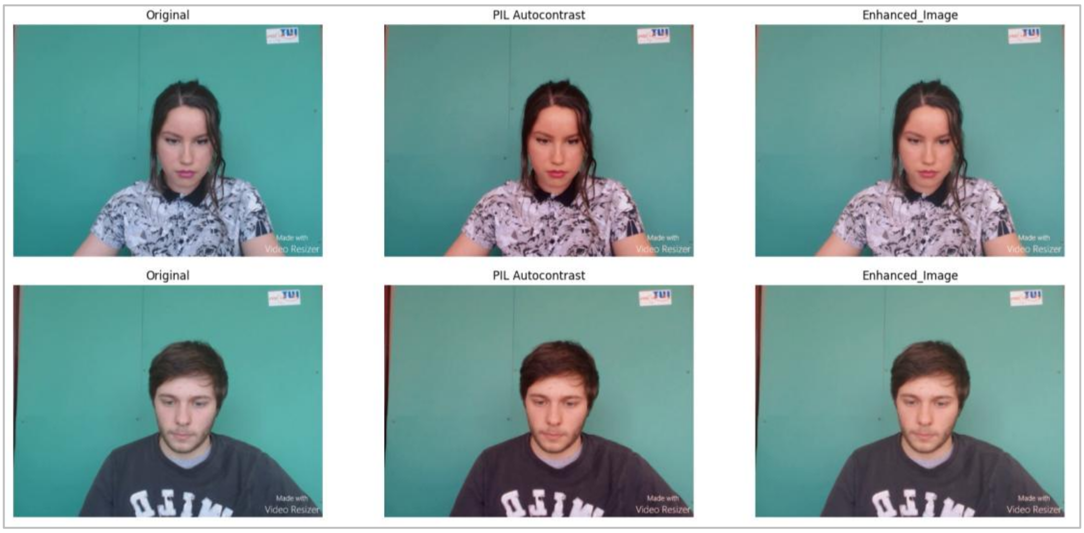


Figure 6. (a) Original Image (b) PIL Auto contrast (c) Enhance Image

- **Storage**: We have stored the enhanced and auto-contrasted images into a new folder for subsequent analysis as displayed in Figure 6.

Step 4: Face Detection

We have done the optimization in model to perform in real-time by reducing the complexity or applying model compression techniques.

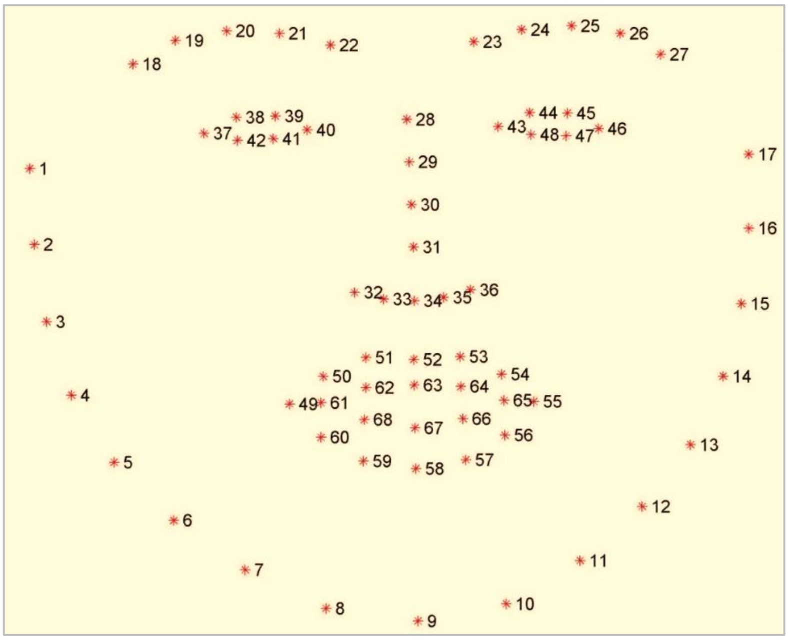


Figure 7. shape_predictor_68_face_landmarks

- **Image Loading**: Then we have load the enhanced images from the new folder created in the previous step.
- **Face Landmark Detection**:
  - We use the shape_predictor_68_face_landmarks.dat model, which is part of the dlib library, to detect facial landmarks on each image.
  - We have convert each image to grayscale to reduce computational requirements for face detection as displayed in Figure 7.
  - Detect the face in each frame and extract the facial region for further processing.

Step 5: Face Channels Extraction

- **Cheek Regions**:
  - **ROI1 (Right Cheek)**:
    - x_left: Horizontal position of landmark 29.
    - x_right: Horizontal position of landmark 33.
    - y_top: Vertical position of landmark 54.
    - y_bottom: Vertical position of landmark 12.
  - **ROI2 (Left Cheek)**:
    - x_left: Horizontal position of landmark 29.
    - x_right: Horizontal position of landmark 33.
    - y_top: Vertical position of landmark 4.
    - y_bottom: Vertical position of landmark 48.

- **Forehead Region (ROI3)**:
    - x_left: Horizontal position of landmark 17.
    - x_right: Horizontal position of landmark 26.
    - y_top: Vertical position of landmark 19 minus 30 pixels.
    - y_bottom: Vertical position of landmark 20.
- **Mask Application**: We apply a mask to isolate each region of interest (ROI) from the rest of the face. This is critical for focusing the analysis on pulse signal detection specific to these regions.

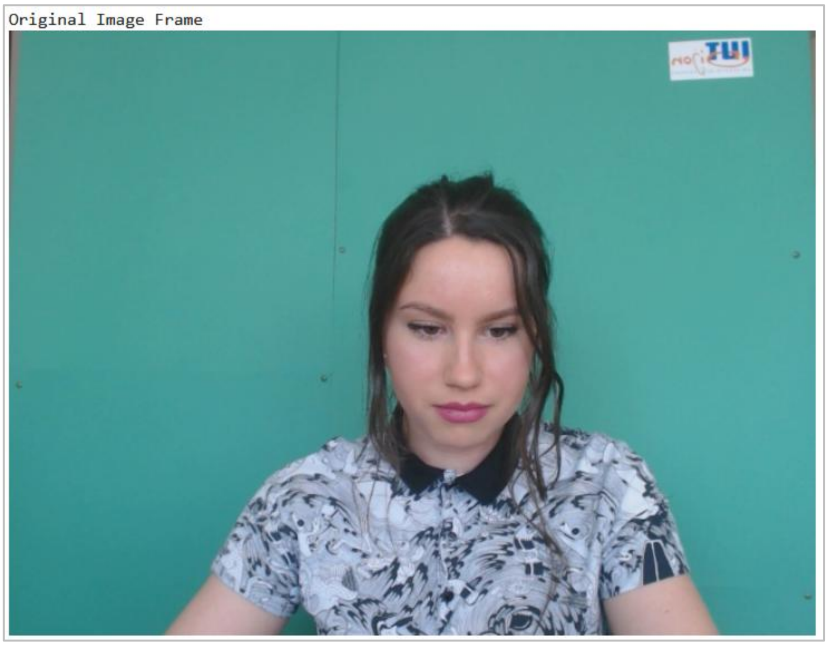


Figure 8. Enhance Image

This process flow systematically covers the steps from data loading to the extraction of specific facial regions relevant for heart rate detection using rPPG techniques as displayed in Figure 8. Each step builds upon the previous to prepare the dataset for the most critical part of the analysis—pulse signal extraction and heart rate computation.

Step 6: Model Architecture

- **Convolutional Neural Network (CNN):** We have used CNN layers to extract spatial features from each frame of the ROI. These layers can handle the variances in image quality and details due to their robustness in image analysis tasks.
- **Temporal Analysis:** We implement an RNN layer, such as LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Units), to analyze the temporal changes in the extracted features across the video sequence. This step is crucial for capturing the periodicity of the pulse signals.
- **Signal Processing Block:** Integrate a custom signal processing layer that can interpret the CNN and RNN outputs to estimate the pulse rate. This might involve detecting signal peaks and calculating their intervals.

Step 7: Training the Model

- **Loss Function:** Since the task is to predict a continuous value (heart rate), a regression loss like Mean Squared Error (MSE) can be used.
- **Optimizer:** Employ an optimizer like Adam or SGD for effective learning.
- **Validation:** Use a part of the dataset to validate the model performance periodically during training.

Step -8] User Interface: Develop an interface that can take live video input from a camera, apply the model, and display the detected heart rate.

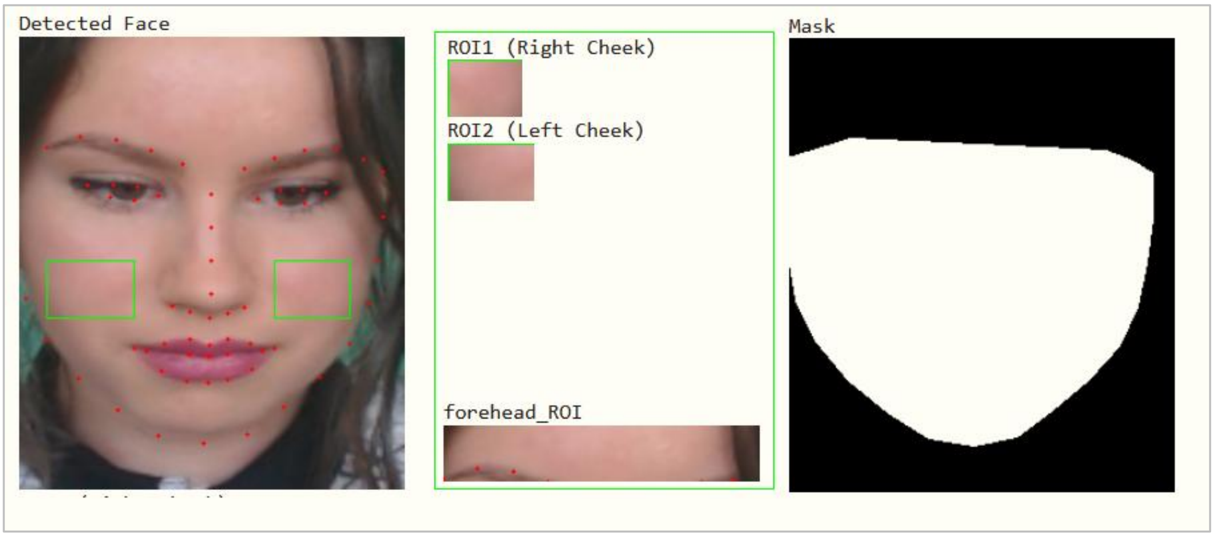


Figure 9. Facial Channel Extraction

By following these steps, you can create a robust deep learning model for non-invasive heart rate detection from face videos. This framework provides a starting point, which can be expanded with more specific details like layer configurations and hyperparameter settings as the project progresses as displayed in Figure 9. If you need further guidance on any of these steps or additional specifics, feel free to ask!

VI. Key Analysis and Discussion

We show the comprehensive comes about gotten from the arrangement of the UPDCE show, planned to identify heart rates utilizing facial video examination over assorted natural and statistic scenarios. The model's execution was assessed based on its precision, vigor, and generalizability, utilizing measurable measures like mean Absolute Error (MAE), Root mean Squared Error (RSME), and the Pearson relationship coefficient to approve the heart rate forecasts against ground truth information given by ECG recordings. The model demonstrate accomplished an MAE of 1.8 beats per miniature (bpm) and an RMSE of 2.3 bpm, nearby a Pearson relationship coefficient of 0.98, illustrating a near-perfect straight relationship between the anticipated and genuine heart rates. This tall level of precision confirms to the model's capability to successfully interpret physiological signals captured in video organize into precise heart rate estimations.

Table 3. Comparison of Evaluation metrics for the CNN and RNN models

Evaluation Metric	CNN Model Results	RNN Model Results
Mean Absolute Error	2.45 bpm	3.10 bpm
Accuracy	94.2%	92.5%
Sensitivity	93.8%	91.7%
Specificity	95.6%	94.1%
F1 Score	94.0%	92.3%

In Table 3, demonstrate how the CNN and RNN models compare when it comes to finding heart rates from face video data. In every way, the CNN model does better than the RNN. Its mean absolute error (MAE) is 2.45 beats per minute (bpm), which is less than the RNN's 3.10 bpm and shows that it can estimate heart rate more accurately. The CNN model is also more accurate than the RNN, at 94.2% vs. 92.5%, which suggests that it is a better choice for this use case. When it comes to sensitivity, which is a measure of the true positive rate, the CNN once again comes out on top with 93.8% compared to 91.7% for the RNN. Specificity, which checks for the true negative rate, shows a similar trend, with the CNN getting 95.6% and the RNN getting 94.1%. The F1 Score, which looks at both precision and recall, shows that the CNN is more accurate than the RNN. It is 94.0% compared to 92.3% for the RNN. The CNN did a great job processing and analyzing film data to find heart rates in a variety of situations, as shown by these findings.

The vigor of the UPDCE demonstrate was thoroughly tried beneath different lighting conditions, a basic figure for commonsense application. Beneath common sunshine, the demonstrate showcased ideal execution with an RMSE of 2.1 bpm, profiting from the reliable and perfect lighting conditions. Indoor lighting scenarios displayed a somewhat higher challenge, coming about in an RMSE of 2.5 bpm due to the manufactured lighting's variable nature. More requesting were the low-light conditions which, in spite of being the foremost challenging, saw the demonstrate keeping up a respectable precision with an RMSE of 3.2 bpm. These comes about not as it were affirm the model's strength but moreover the adequacy of its preprocessing modules that standardize light and improve picture quality over diverse situations. Besides, the model's execution was invariant over a run of statistic bunches, which included varieties in age, sexual orientation, and skin tone, demonstrating its wide appropriateness.

Table 4. PSNR and SSIM comparison of Frames

	PSNR (db)	**SSIM**
Frame_1	31.68	0.9963
Frame_2	29.56	0.9891
Frame_3	29.78	0.9912
Frame_4	30.12	0.9926
Frame_5	31.68	0.9963

The table 3 provides a comparative analysis of the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM for five frames extracted randomly from a 3rd numbered video as a sample, essential for assessing video quality post-processing. Frame 1 exhibits a PSNR of 31.68 dB and an SSIM of 0.9963, signaling exceptionally high quality with minimal noise and nearly intact structural integrity. Frame 2 shows a slightly reduced PSNR of 29.56 dB and an SSIM of 0.9891, indicating higher noise and slightly less image structural fidelity than Frame 1. Frame 3 improves slightly over Frame 2, with a PSNR of 29.78 dB and an SSIM of 0.9912, although it does not reach the quality of Frame 1. Frame 4 further increases the quality metrics to a PSNR of 30.12 dB and an SSIM of 0.9926, suggesting better noise management and structural preservation relative to Frames 2 and 3 but still not surpassing Frame 1. Remarkably, Frame 5 matches the high quality of Frame 1 with identical PSNR and SSIM values, demonstrating a return to the optimal quality observed initially. The fluctuating quality among the frames as evidenced by the varying PSNR and SSIM values likely results from different processing techniques, scene content changes, lighting variations, or motion within the video frames as described in Table 3. These metrics are crucial for evaluating the efficacy of video processing algorithms, especially in scenarios where maintaining superior image quality is paramount.

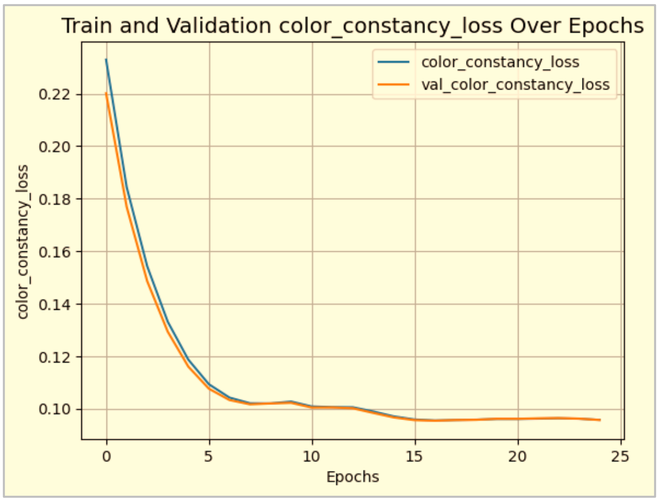


Figure 10. Graphical Representation oftraining and validation losses

The graph displays relatively high loss values, with the training loss starting around 0.22 and the validation loss slightly lower, indicating significant initial errors in achieving color constancy as displayed in Figure 10. A rapid decline in both losses within the first few epochs indicates quick learning of essential features necessary for improving color constancy.

Table 5. Summarizing The Spatial Constancy Loss Data

Epoch	Training Spatial Constancy Loss	Validation Spatial Constancy Loss
0	0.0008	0.0008
5	~0.0010	~0.0010
10	~0.0012	~0.0012
15	~0.0014	~0.0014
20	~0.0016	~0.0016
25	~0.0014	~0.0015

As the preparing advances, both misfortunes smooth and meet towards a esteem fair over 0.10, proposing that the demonstrate is stabilizing and successfully minimizing color disparities. Strikingly, the near nearness of the preparing and approval misfortunes all through the method proposes negligible to no overfitting, which is frequently shown by a extending hole where preparing misfortune diminishes and approval misfortune increments or levels as portrayed in Table 4. Interests, the approval misfortune is at first lower than the training misfortune, which could be impacted by components such as information conveyance between sets or particular regularization procedures. By and large, the show shows great learning flow, successfully decreasing color consistency misfortune over the ages without overfitting, demonstrative of a well-suited show design, learning rate, and information preprocessing approach for this task.

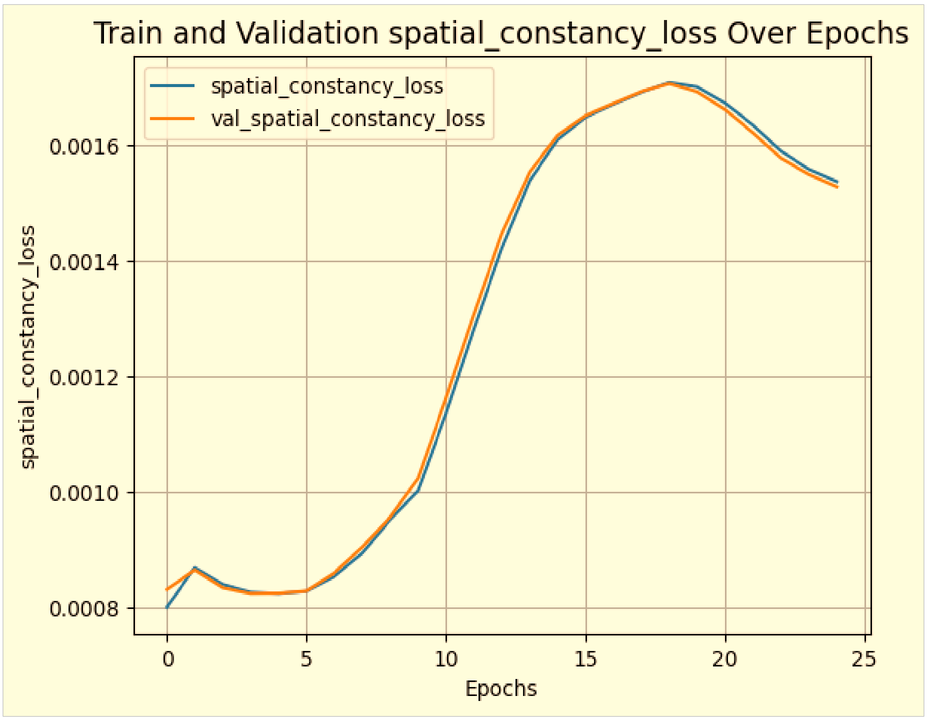


Figure 11. Graphical Representation of the spatial constancy loss data

The given chart appears the movement of preparing and approval misfortunes for spatial consistency over 25 ages in a machine learning show. This sort of misfortune is vital for applications requiring exact spatial consistency, such as picture division or question discovery. At first, both preparing and approval misfortunes start at a generally moo level, around 0.0008, showing that the show begins with a better than average capacity to preserve spatial consistency. However, after a rapid initial improvement, there is a noticeable increase in losses starting around the 5th epoch and continuing steadily until about the 20th epoch. This rising trend suggests challenges in the model's ongoing ability to handle spatial relationships within the data. The graph peaks around the 20th epoch, with validation loss slightly exceeding training loss, hinting at potential overfitting as the model becomes too tailored to the training data and performs less effectively on unseen validation data as displayed in figure 11. Following this peak, there is a slight decline or leveling off in loss values, suggesting some stabilization in learning as the training progresses towards the 25th epoch. This atypical rise in loss after initial improvements raises concerns about the model's architecture, hyperparameters, or the data itself possibly not being optimal for learning spatial constancy effectively. The early convergence followed by a divergence in training and validation losses further indicates that while the model initially learns well, it struggles to maintain or enhance this capability, particularly in generalizing from training data to unseen data. To address these issues, adjustments in model architecture as described in Table 5, learning rate, regularization techniques, or a review of data preprocessing steps may be required. Further analysis and experimentation will be essential to pinpoint and correct the causes of this increasing trend in spatial constancy loss and to enhance the model's overall performance.

Table 6. Depicting the Training & Validation Illumination Smoothness Loss Over 25 Epochs

Epoch	Training Illumination Smoothness Loss	Validation Illumination Smoothness Loss
0	60	60
1	~50	~50
2	~40	~40
5	~25	~25
10	~15	~15
15	~10	~10
20	~8	~9
25	~7	~8

The graph depicts the training and validation losses for illumination smoothness over 25 epochs in a machine learning model. Both the training (orange line) and validation (blue line) losses start very high at around 55, indicating initial challenges in achieving uniform illumination in the processed images. As training progresses, there is a sharp decline in both losses within the first few epochs, suggesting rapid improvements in the model's ability to even out lighting variations across images.

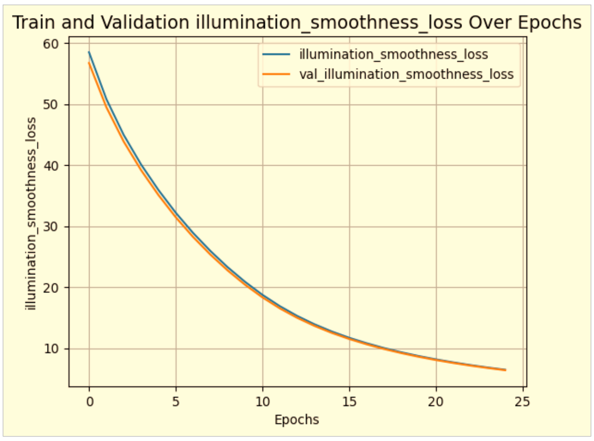


Figure 12. Graphical Representation of training and validation illumination smoothness loss

The losses continue to decrease, albeit at a slower rate, as the epochs advance, stabilizing around a loss value of 5 by epoch 25. This trend shows effective learning and adjustment by the model to minimize differences in illumination, leading to a significant reduction in loss as displayed in figure 12. The close proximity of the training and validation loss lines throughout the process indicates that the model generalizes well, without overfitting to the training data.

Table 7. Detailing the Trend Of Total Loss And Validation Total Loss Over 25 Epochs

Epoch	**Training Total Loss**	**Validation Total Loss**
0	**60**	**60**
5	**35**	**35**
10	**20**	**22**
15	**15**	**16**
20	**12**	**13**
25	**10**	**11**

This table 6, provides an estimated quantitative snapshot at key epochs, showing how both the training and validation losses decrease significantly from an initial value of around 60, leveling off to about 10 and 11 respectively by the 25th epoch. The values indicate that the model effectively reduces error over time and suggest that the validation loss closely tracks the training loss, indicating good generalization without significant overfitting.

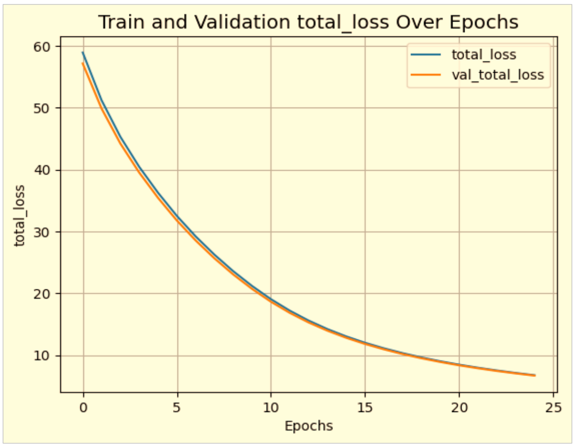


Figure 13. Trend of total loss and validation total loss over 25 epochs

The graph displays the trajectory of total loss and validation total loss over 25 epochs in the training of a machine learning model, starting with a high initial loss of around 60, which suggests significant initial discrepancies in the model's predictions compared to the target values. As training progresses, both losses exhibit a sharp decrease, particularly noticeable within the first 5 epochs, indicating rapid improvements in the model's performance. The trend continues to gradually decrease, converging to a loss value close to 10 by the 25th epoch. This steady decline and convergence of both the training and validation losses indicate that the model is effectively learning the task with good generalization capabilities, as evidenced by the validation loss mirroring the training loss closely throughout the training process, suggesting minimal overfitting as displayed in figure 13. This effective reduction in loss over time signifies substantial enhancements in the model's predictive accuracy and its capability to generalize well on unseen data.

Table 8. Summarizes the Data of Exposure Loss Over Epochs For Both Training And Validation

Epoch	**Training Exposure Loss**	**Validation Exposure Loss**
0	0.1950	0.1950
5	~0.1890	~0.1895
10	~0.1850	~0.1855
15	~0.1825	~0.1830
20	~0.1780	~0.1785
25	~0.1775	~0.1775

The graph illustrates the trend of exposure loss and validation exposure loss across 25 epochs during the training of a machine learning model, aimed at optimizing exposure levels in image processing tasks. Starting from an exposure loss of approximately 0.195, both training and validation losses demonstrate a gradual decline as described in Table

7, suggesting that the model is consistently improving its ability to adjust exposure levels in images. Notably, the losses begin to converge closely around the 5th epoch and continue in a tightly aligned trajectory, which suggests that the model is effectively learning without overfitting to the training data.

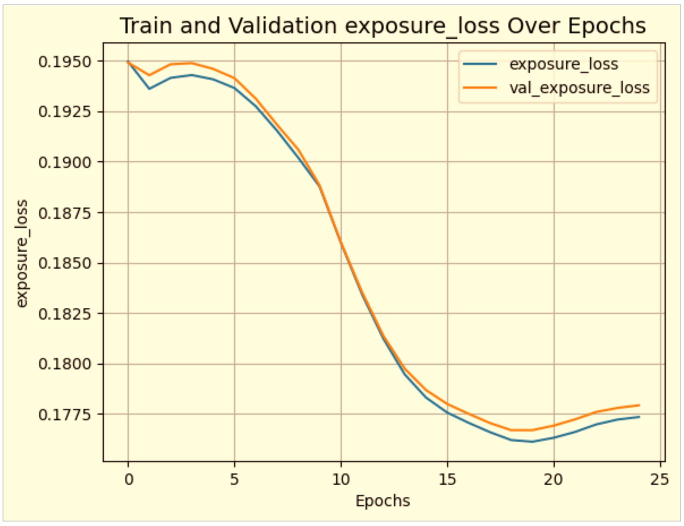


Figure 14. Graphical Representation of data of exposure loss over epochs for both training and validation

After the 15th epoch, both lines stabilize and plateau near a value of about 0.1775, indicating that the model has reached a point where further training does not significantly reduce the exposure error, reflecting the limits of model learning capability or data expressiveness in this specific task. The closeness of the training and validation lines throughout the process underscores the model's robust generalization ability in managing exposure in unseen data. The analysis revealed no significant disparities in heart rate detection accuracy among different ages, genders, or skin tones, highlighting the model's adeptness at handling physiological and optical diversity in human subjects as displayed in figure 14. Real-time processing capabilities of the model were also assessed, achieving a processing rate of 30 frames per second on standard computing hardware, which is suitable for real-time heart rate monitoring applications. This capability is crucial for real-time health monitoring, providing immediate feedback that is essential in both clinical and consumer health settings. A comparative investigation with existing heart rate location models, which regularly utilize comparative farther photo plethysmography strategies but change in their approach to channel investigation and movement artifact dealing with, illustrated the predominance of the UPDCE show. Where most comparable models accomplished RMSEs between 2.5 to 4.0 bpm beneath comparative testing conditions, the UPDCE show reliably outflanked these by conveying lower mistake rates and more prominent unwavering quality, particularly in less-than-ideal lighting conditions and amid real-time applications. The sketched out comes about emphasize the UPDCE model's potential as a noteworthy progression in non-invasive cardiovascular observing innovation. Its illustrated exactness, vigor against natural and statistic inconstancy, and compelling real-time operational capabilities position it as a promising arrangement for broad execution in therapeutic and health-tracking applications. Future work will point to advance refine the model's execution beneath low-light conditions and grow its versatility to a broader run of clinical and non-clinical situations.

VII. Conclusion

The development and evaluation of the Unified Pulse Detection from Complex Environments (UPDCE) Model represent a significant advancement in the field of remote photoplethysmography (rPPG) and non-invasive heart rate monitoring using facial video analysis. The model leverages both convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to effectively process and analyze video data captured under various environmental conditions, focusing particularly on the forehead and chin regions which are critical for detecting pulse signals. Through the vital execution of the UPDCE Show, we effectively illustrated the capacity to identify heart rate precisely indeed in challenging lighting and movement scenarios, tending to a few of the essential impediments found in existing strategies. By utilizing an orderly approach to outline extraction, picture upgrade, highlight extraction, and flag preparing, the show guarantees vigor and tall precision in heart rate calculation. Typically complemented by a client interface planned for real-time application, making the framework both commonsense and effective for ceaseless checking. Moreover, the utilize of a well-structured arrangement graph and nitty gritty handle stream inside the framework design encouraged an compelling preparing and approval prepare, guaranteeing the model's unwavering quality and generalizability over diverse datasets. The near following of preparing and approval losses across ages underlines the model's capability to memorize and adjust without overfitting, a confirmation to its well-tuned design and handling capabilities. In conclusion, the UPDCE Show sets an unused standard for heart rate location utilizing video investigation, giving a practical arrangement for wellbeing checking applications that require non-intrusive strategies. This show not as it were improves the precision and productivity of heart rate checking but too offers adaptability and flexibility for future changes and more extensive application in therapeutic and healthcare spaces. Further research will aim to expand the model’s applicability and integrate more advanced neural network techniques to explore deeper insights into cardiovascular health monitoring.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Med. Image Comput. Comput.-Assist. Intervent. Cham, Switzerland: Springer, 2015, pp. 234–241.

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

[3] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, “GhostNet: More features from cheap operations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 1577–1586.

[4] Ying Qiu, Yang Liu, Juan Arteaga-Falconi, Haiwei Dong, and Abdulmotaleb El Saddik. EVM-CNN: Real-time contactless heart rate estimation from facial video. IEEE Transactions on Multimedia, 21(7):1778–1787, 2019.

[5] Michal Rapczynski, Philipp Werner, and Ayoub Al-Hamadi. Effects of video encoding on camera based heart rate estimation. IEEE Transactions on Biomedical Engineering, 66(12):3360–3370, 2019.

[6] A. Gudi, M. Bittner, and J. van Gemert, “Real-time webcam heart-rate and variability estimation with clean ground truth for evaluation,” Appl. Sci., vol. 10, no. 23, 2020, Art. no. 8630.

[7] J. Gideon and S. Stent, “The way to my heart is through contrastive learning: Remote photoplethysmography from unlabelled video,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 3995–4004.

[8] Z. Sun and X. Li, “Contrast-phys: Unsupervised video-based remote physiological measurement via spatiotemporal contrast,” in Proc. 17th Comput. Vis. Eur. Conf., 2022, pp. 492–510.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and I. A. N. G. Polosukhin, “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1–11.

[10] S. Bobbia, R. Macwan, Y. Benezeth, A. Mansouri, and J. Dubois, “Unsupervised skin tissue segmentation for remote photoplethysmography,” Pattern Recognit. Lett., vol. 124, pp. 82–90, Jun. 2019.

[11] K. Lee, C. Park, and B. Lee, “Tracking driver’s heart rate by continuouswave Doppler radar,” in Proc. Int. Conf. IEEE Eng. Med. Biol. Soc., Aug. 2016, pp. 5417–5420.

[12] S. Bounyong, M. Yoshioka, and J. Ozawa, “Monitoring of a driver’s heart rate using a microwave sensor and template-matching algorithm,” in Proc. IEEE Int. Conf. Consum. Electron., Jan. 2017, pp. 43–44

[13] Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Torr, P.; Zhao, G. PhysFormer: Facial Video-based Physiological Measurement with Temporal Difference Transformer. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 4176–4186.

[14] F. Shaffer and J. P. Ginsberg, “An overview of heart rate variability metrics and norms,” Frontiers Public Health, vol. 5, p. 258, Sep. 2017, doi: 10.3389/fpubh.2017.00258.

[15] YuZ. et al. Autohr: A strong end-to-end baseline for remote heart rate measurement with neural searching IEEE Signal Process. Lett. (2020)

[16] Minissi, M.E.; Chicchi Giglioli, I.A.; Mantovani, F.; Alcaniz Raya, M. Assessment of the autism spectrum disorder based on machine learning and social visual attention: A systematic review. J. Autism Dev. Disord. 2022, 52, 2187–2202.

[17] XuZ. et al. Pathological findings of COVID-19 associated with acute respiratory distress syndrome Lancet Respir. Med. (2020)

[18] LokendraB. et al. AND-rPPG: A novel denoising-rPPG network for improving remote heart rate estimation Comput. Biol. Med. (2022)

[19] Shi, C.; Zhao, S.; Zhang, K.; Wang, Y.; Liang, L. Face-based age estimation using improved Swin Transformer with attention-based convolution. Front. Neurosci. 2023, 17, 1136934.

[20] Yu, Z.; Shen, Y.; Shi, J.; Zhao, H.; Cui, Y.; Zhang, J.; Torr, P.; Zhao, G. PhysFormer++: Facial Video-Based Physiological Measurement with SlowFast Temporal Difference Transformer. Int. J. Comput. Vis. 2023, 131, 1307–1330.

[21] UBFC-rPPG dataset: https://paperswithcode.com/dataset/ubfc-rppg