Smartphone-Based Facial Video Analysis for Real-Time Vital Sign Estimation and User Engagement

Jonathan B. Kierce and Aydan A. Kumar

Monash Department of Electrical Engineering

Abstract

Remote photoplethysmography (rPPG) on commodity smartphones promises contactless cardiovascular monitoring, yet practical on-device implementations and real-time validation remain limited. We present a fully on-device iOS pipeline that locates a facial forehead region of interest, applies CHROM projection to the red-green-blue signal, band-pass filters the signal, and performs peak detection to derive heart rate (HR) and root mean square of the successive differences-based heart rate variability (HRV). Furthermore, an 11-feature logistic-regression model estimates the atrial fibrillation (AF) probability of the sample. Application performance in HR and HRV calculation was validated against a Polar H10 ECG through 30-second recordings across stillness, paced breathing, and head-nod/shake tasks. The application performed strongly in measuring heart rate, achieving a mean average error (MAE) of 1.48 beats per minute. However, the HRV calculation suffered from noise in the rPPG signal and achieved an MAE of 63.5 milliseconds. The AF classifier, trained on a publicly available PPG dataset, achieved 94.2% accuracy but lacks end-to-end rPPG validation, highlighting the need for labelled rPPG AF data. Overall, the results demonstrate accurate AF classification and on-device HR measurement at rest, identify limitations for short-window HRV, and suggest protocol refinements for robust real-world use.

1 Introduction

In recent years, there has been growing demand for accessible and contactless methods to monitor cardiovascular health. The widespread availability of smartphones equipped with high-resolution cameras and advanced processing capabilities has created new opportunities for mobile health applications. Remote photoplethysmography (rPPG) leverages these capabilities to measure vital signs, such as heart rate (HR), by analysing subtle colour changes in the skin caused by blood flow (Wang et al., 2017). Compared to traditional contact-based methods, mobile rPPG offers a scalable, low-cost, and user-friendly solution that enables continuous and unobtrusive monitoring in real-world environments.

Atrial fibrillation (AF) is the most common type of heart arrhythmia, a condition in which the heart beats irregularly. It is thought to affect one in every three to five individuals over the age of 45 (Linz et al., 2024). AF is caused by abnormal electrical activity in the atria (upper chambers) of the heart and is typically faster than a healthy heartbeat. It is the leading cardiac cause of stroke. Risk factors for AF include high blood pressure, underlying heart or lung disease, and endocrine disorders such as diabetes (Nesheiwat et al., 2023).

HR variability (HRV) is a measure of the variation in time intervals between heartbeats. HRV is closely related to the autonomous nervous system (ANS), blood pressure, and mental well-being. The ANS is responsible for regulating HR, blood pressure, breathing, and digestion, among other key tasks (Li et al., 2019).

Currently, the main methods of monitoring AF and HRV are mobile contact-based electrocardiogram (ECG) and phone–based photoplethysmography (PPG). However, both of these are contact-based methods. ECG requires the attachment of electrodes to the body, which can cause discomfort for the user. Similarly, mobile-based PPG methods require users to either place their finger on the phone camera or rely on external sensors, which can feel awkward and discourage frequent use (Li et al., 2019) (Odinaev et al., 2023). In recent years, there has been a growing interest in estimating HR vital signs using rPPG which involves using a camera to detect changes in skin colour that occur due to blood flow.

While the overall rPPG process for detecting AF and HRV is consistent across studies, implementations often modify parts of the pipeline to improve performance. Video is captured using a camera, then colour signals are extracted and amplified using various techniques to detect a heart pulse. This extracted signal is later passed into a machine learning (ML) model or algorithm to extract vital sign signals. To extract colour data from the video, a region of interest (ROI) is obtained, and the pixels within the ROI are extracted and averaged in each frame to get a

time series of RGB pixel values (Odinaev et al., 2023). Popular selections for the ROI are the forehead, nose, cheeks, and lips (Li et al., 2024). These colour signals are then amplified to extract the PPG signal. Many methods have been explored for this step. Some of the best-performing methods are Spatial Subspace Rotation (SSR), the CHROM algorithm, orthogonal matrix image transformation (OMIT), local group invariance (LGI), and plant-orthogonal-to-skin (POS). The simplest approach is to use the green colour signal (Odinaev et al., 2023) (Haugg et al., 2022). It is common to pass the heart signal through a type of filter which is used to eliminate any impossible heart rates. Most implementations use a bandpass filter. These filtered HR signals are then passed into an ML model to detect AF or an algorithm to detect HRV. AF detection has been done with simple ML models such as support vector machines (SVMs), random forests, and boosted trees; however, more complex networks, including convolutional neural networks (CNNs), have had better results (Sun et al., 2022). HRV can be expressed using various metrics, which are calculated by obtaining the inter-beat intervals from the data. Some common metrics are the root mean square of the successive differences (RMSSD), the standard deviation of the NN intervals (SDNN), and the proportion of the number of pairs of successive beat-to-beat intervals that differ by more than 50 milliseconds (pNN50). Each of these metrics can be used to get insights into different aspects of an individual's health (Li et al., 2019) (Odinaev et al., 2023).

Previous research has shown that while rPPG-based systems have made significant advancements and perform well, the field still has notable gaps. Many existing mobile-based applications for detecting AF and HRV rely on traditional PPG methods, which require direct physical contact with a device. This presents a challenge in environments where minimising cross-contamination is essential, such as hospitals. In contrast, the non-contact nature of rPPG offers a more accessible, comfortable, and hygienic alternative, reducing both user discomfort and the risk of contamination. Furthermore, rPPG-based systems have the advantage of requiring no extra equipment, making them both accessible and easy to use.

However, despite the promise of rPPG, its integration into mobile platforms remains limited. Much of the existing research focuses on controlled laboratory setups using specialised equipment, with little emphasis on real-world usability or implementation on consumer-grade mobile devices (Liu et al., 2020). Furthermore, many studies that utilise mobile hardware still rely on post-processing, where data captured on a smartphone is transferred to a computer or cloud environment for analysis (Huang and Dung, 2016). This separation between acquisition and processing limits the practicality and immediacy of rPPG-based monitoring in real-world applications. In addition, challenges such as variable lighting conditions, motion artefacts, and differences in camera hardware continue to affect the accuracy and reliability of current solutions (Wang et al., 2017). These constraints highlight a clear gap in the development of robust, accessible, and mobile-friendly rPPG systems that can be used in real situations. Addressing these limitations is essential to make non-contact cardiovascular monitoring practical, scalable, and suitable for widespread adoption in both clinical and everyday contexts.

This paper makes a significant contribution to the field by presenting a fully integrated iOS application that:

(a) **measures heart rate;**

(b) **measures heart rate variability;**

(c) **detects atrial fibrillation; and**

(d) ***performs (a)–(c) in real time on-device.***

By performing all processing locally, the system demonstrates that accurate, real-time rPPG analysis and AF detection can be achieved on consumer-grade hardware without reliance on external computing resources.

2 Methodology

2.1 rPPG Signal Extraction

Our rPPG pipeline begins by configuring the front-facing camera to stream frames with timestamps. The session is tuned for a high, stable frame rate and auto-exposure, with white balance enabled. Once a face is reliably detected, we freeze both these settings, preventing drifts in brightness that could negatively impact the extraction of the subtle colour changes rPPG relies on.

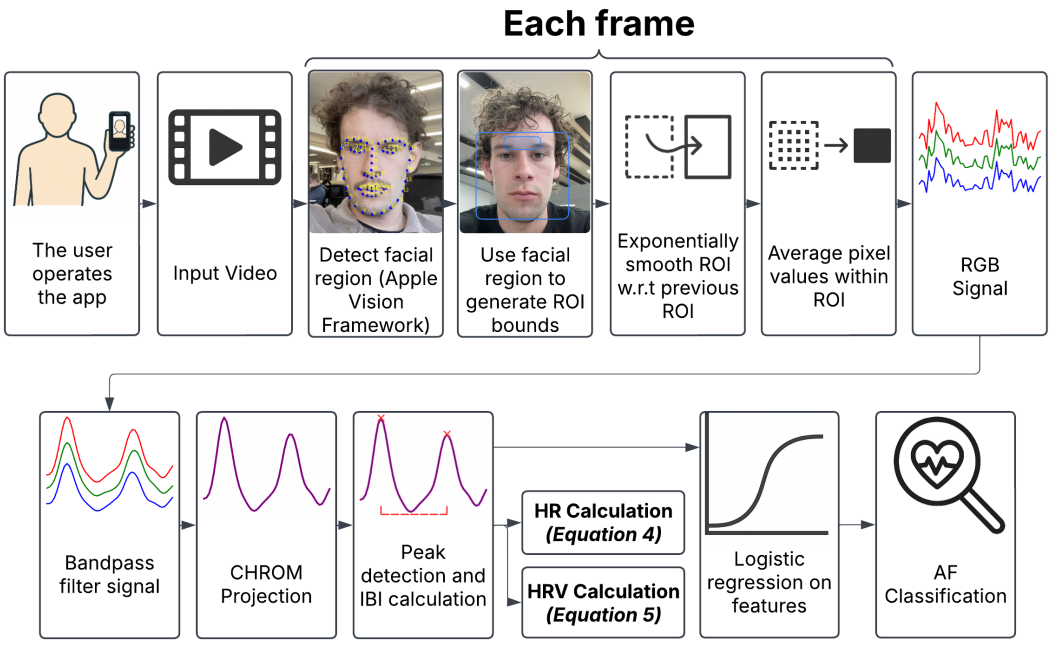

Figure 1: Per-frame rPPG processing pipeline used in this study. A face detector defines a smoothed ROI from which mean RGB values are extracted, band-pass filtered, and CHROM-projected; peak detection yields IBIs for HR (Eq. (4)) and HRV (Eq. (5)), and derived features feed a logistic-regression AF classifier.

Each frame is analysed using Apple's Vision (Apple Inc., 2025b) framework to locate the user's facial bounding box and then the forehead area. We trim and smooth the raw face bounding box to form a stable ROI that follows the subject while reducing jitter.

Within the ROI, we average the red, green, and blue (RGB) channels of every pixel. Per-frame timestamps are maintained, which allows the application to recover from dips in frame processing rates.

We then apply the CHROM projection as suggested by de Haan and Jeanne (2013) to decouple pulse-induced chrominance changes from motion artefacts. After normalising the mean RGB traces within the region of interest, let $\tilde{r}(t)$, $\tilde{g}(t)$, and $\tilde{b}(t)$ denote the temporal colour fluctuations. The chrominance components are

$$X(t) = 3\,\tilde{r}(t) - 2\,\tilde{g}(t), \tag{1}$$

$$Y(t) = \tfrac{3}{2}\,\tilde{r}(t) + \tilde{g}(t) - \tfrac{3}{2}\,\tilde{b}(t), \tag{2}$$

with corresponding standard deviations σ_X and σ_Y evaluated over the sliding analysis window. The final projected pulse trace is then

$$C(t) = X(t) - \frac{\sigma_X}{\sigma_Y}\,Y(t). \tag{3}$$

The idea of this projection is to emphasise the green-dominated pulse component, while suppressing motion-correlated variations that affect all RGB channels similarly.

The projected signal passes through a fourth-order filter suggested by Butterworth (1930) with low and high pass frequencies of 0.7 and 4Hz respectively. This removes slow illumination drift while retaining the cardiac component expected between 42 and 240 beats per minute. Peak candidates are then selected by identifying local maxima that satisfy a 0.3s refractory period, suppressing duplicate detections within the same cardiac cycle.

With the location of the peaks, HR is calculated by averaging the inter-beat intervals $\Delta t_i = t_i - t_{i-1}$ between successive peaks t_i:

$$\mathrm{HR} = \frac{60(N-1)}{\sum_{i=2}^{N} \Delta t_i} \tag{4}$$

and HRV with RMSSD, which captures short-term variability in the peak-to-peak spacing:

$$\mathrm{HRV}_{\mathrm{RMSSD}} = \sqrt{\frac{1}{N-2}\sum_{i=2}^{N-1}(\Delta t_{i+1} - \Delta t_i)^2} \tag{5}$$

2.2 AF Classification

After the rPPG signal is extracted, a set of 11 cardiovascular features are derived. These features are HR, HR mean, beats per second, number of detected beats, mean RR interval, median RR interval, SDNN, RMSSD, PNN50, and statistical descriptors of the PPG signal (mean and standard deviation). The features are then passed into a logistic regression model to estimate the likelihood that a given sample exhibits AF. Logistic regression models the log-odds of AF occurrence as a linear combination of these standardised features, applying the sigmoid function to convert this into a probability between 0 and 1.

Model development was based on the MIMIC PERform AF dataset (Charlton, 2022), which contains synchronised PPG and ECG recordings from 35 ICU patients, each with 20 minutes of continuous physiological data. Of these, 19 patients exhibited AF during the recording period, while 16 were labelled as non-AF. Due to the limited availability of open-access AF datasets, this collection served as the primary benchmark for all model development.

Initial experiments explored a range of machine learning approaches, from traditional classifiers such as SVMs and decision trees to more complex deep learning architectures, including CNNs and recurrent neural networks. However, these models consistently failed to converge effectively; training and validation performance plateaued early, suggesting insufficient data diversity and overfitting to patient-specific characteristics. Consequently, a shift towards a feature-based classification approach was pursued. Logistic regression was ultimately selected as the most practical and reliable solution, offering strong interpretability, reduced overfitting risk, and minimal computational requirements suitable for mobile implementation.

To maximise the available training data, each 20-minute recording was segmented into 30-second windows with a 5-second rolling overlap. The resulting snippets were filtered using the same bandpass filter described in Section 2.1 to remove baseline drift and high-frequency noise, followed by normalisation to ensure consistent input scaling. The dataset was split at the patient level into 75/15/15 training, validation, and test subsets to prevent data leakage and ensure evaluation on unseen subjects. For the logistic regression model specifically, a standard 80/20 train–test split was applied. The final model achieved an accuracy of 94.2% on the test set.

2.3 App User Experience

The app has three main views. The first view is the main collection view, and is shown when the user opens the application. The second view is the measurement screen, which is shown when the user completes a measurement, so they can review their heartbeat waveform and look at their HR, HRV and AF results. The third screen is a simple settings screen, including functionality to use 30 or the maximum frame rate available, and setup for networking to a computer as described in section 2.4. App features can be divided into the following main subsections:

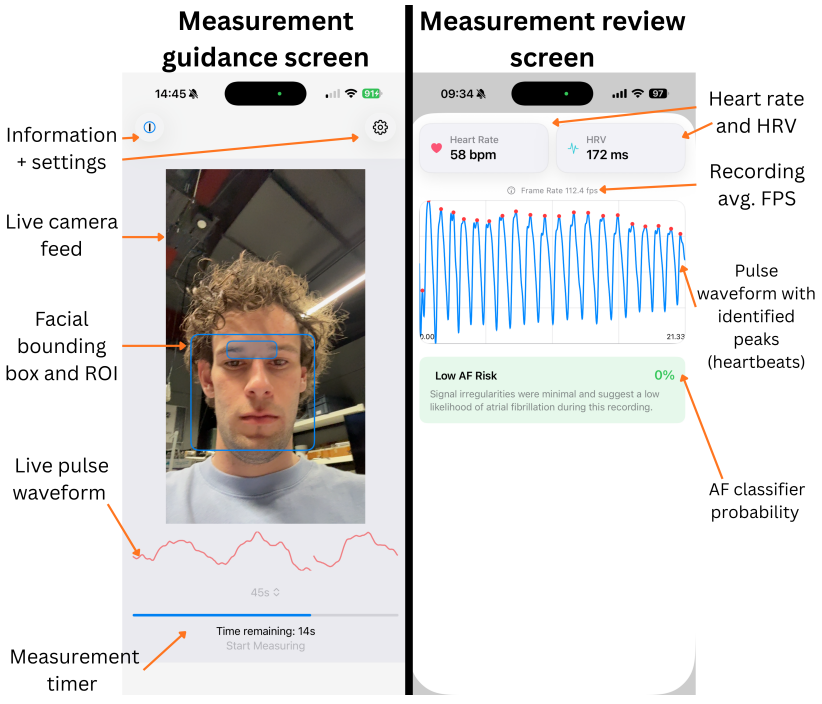

Figure 2: The two important screens in the app. The measurement guidance screen has features to assist the user in taking an accurate measurement, while the measurement review screen allows for quick review of the measurement waveform and derived statistics.

2.3.1 Measurement recording assistance features

In order to assist the user with taking a good rPPG reading, the app provides several useful features. Firstly, the live visual display, with a live bounding box plotted on top, helps the user to ensure their face is correctly positioned and that the face is being correctly located in the frame. Secondly, the user has the option to choose a measurement duration (15–90 s), allowing flexibility between shorter, convenient readings and longer sessions that yield more robust rPPG estimates.

Once the user begins a measurement by pressing "Begin Measurement", two further features are activated. Firstly, a live heartbeat waveform is shown at the bottom

of the screen. This provides instant feedback to the user, allowing them to check the quality of the signal during the recording process. If they can identify a heartbeat, the measurement is likely to be accurate, whereas if one is not visible, the conditions may need to be altered for the best measurement. Secondly, a function runs using the iPhone's accelerometer and the bounding box data. If either the accelerometer readings or the change in bounding box location is too large, the screen turns red, indicating the user is moving too much. This provides real-time feedback.

2.3.2 Measurement analysis and review features

Once a measurement is complete, a screen appears that demonstrates the recording, including waveform, calculated HR, HRV, and AF classification. The user then has the ability to review the waveform, which can be pinched to zoom or scrolled left and right. Peaks are identified and, if any outlier peaks have been detected, the user is given the option to run calculations including or excluding them. Outlier beats are detected by examining consecutive inter-beat intervals (IBIs). Any interval shorter than 60% of the median IBI is treated as a potential double-beat artifact; if adding the next interval yields a combined duration between 80% and 120% of the median, the pair is merged and the second peak is flagged as an outlier. We chose to keep outliers visible because the app includes AF detection, while these beats would normally be discarded for healthy users, we cannot make that assumption for users with irregular rhythms.

2.3.3 Other features

Other features include an informational screen, accessible by the *I* icon in the top left. This contains a best-practice guide for getting the best rPPG signal. Settings are also available by clicking the cog icon in the top right corner. This gives the user the ability to control certain settings, such as saving the video clip after a measurement, or to choose between 30 FPS or the maximum available frame rate when streaming.

2.4 Validation Methodology

Typically, validation of an rPPG product takes place by benchmarking against one of the datasets out there, such as the VIPL-HR-V2 Database (Niu et al., 2018). In the case of our app, benchmarking against a pre-recorded dataset is not useful, as key features of our app, such as locking exposure and ISO on the camera, along with visual real-time guidance for the user, are designed to improve rPPG performance. We therefore validated the system in real-time by pairing the app with an external ECG reference.

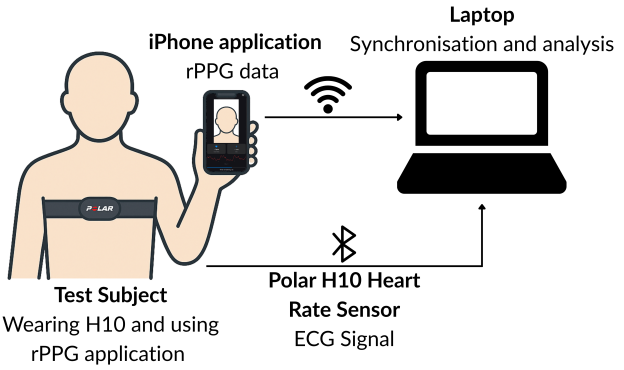

Figure 3: Experimental setup: iPhone rPPG and Polar H10 (130 Hz ECG via BLE) stream to a laptop for synchronisation and analysis.

During each measurement, the iOS app streams JSON summaries to a laptop-based recorder over a WebSocket connection. The payloads include synchronised timestamps, raw and filtered CHROM waveforms, peak annotations, and the HR and HRV estimates produced on-device, accompanied by telemetry on frame cadence and streaming health. In parallel, the recorder connects to a Polar H10 chest strap over Bluetooth, capturing a continuous 130 Hz ECG trace. The H10 was chosen as it is relatively inexpensive compared to full ECG systems, but provides good accuracy compared to a comparable PPG system. Schaffarczyk et al. (2022) found the device to have R-peak accuracy under 1ms when compared to a full ECG, which is more than accurate enough for rPPG validation. At the end of the capture, the recorder packages the synchronised rPPG and ECG signals together with metadata describing the session, which allows the analysis scripts to resample both streams, derive reference HR and HRV metrics from the ECG according to equations (4) and (5), and compare them directly against the app's estimates.

10 subjects were used for validation, comprised of 7 male and 3 female participants (age 30.6 ± 14.9 years; range 18–54). Skin tone was self-reported using the Fitzpatrick phototype scale (types I–VI) (Fitzpatrick, 1988) with I:1, II:3, III:1, IV:4, V:0, VI:1. No participants reported cardiovascular conditions. Each participant completed eight 30 s measurements: (i) six standard app-use recordings; (ii) one paced-breathing trial (6 s inhale, 4 s exhale) to reduce HR; and (iii) one yes/no Q&A trial to elicit head motion. Recordings were excluded if the ECG stream dropped out, the app stream dropped out, or the participant terminated the task. Recordings were discarded if the ECG stream dropped out, the app stream dropped out, or if the participant stopped for any reason.

3 Results

3.1 App Validation Results

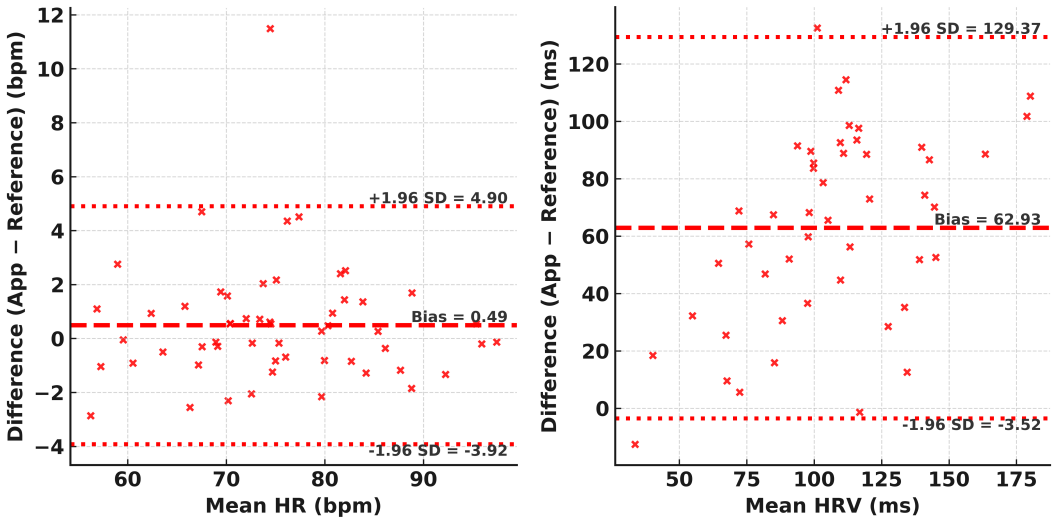

Figure 4: HR error by condition type. 30-second tests were conducted on 10 subjects

Agreement analysis (Bland–Altman). Figure 4 shows Bland–Altman plots for HR and HRV using *Still* recordings (HR: $n = 55$; HRV: $n = 48$).

Heart rate: The mean difference (App–H10) was $\bar{d} = 0.49$ bpm with 95% limits of agreement $[-3.92, 4.90]$ bpm (LoA 95% CIs: lower $[-5.32, -2.50]$, upper $[2.85, 7.21]$). MAE = 1.48 bpm, RMSE = 2.28 bpm, and 98.2% of estimates were within ±5 bpm of H10. Lin's concordance correlation coefficient (CCC) was 0.974.

heart rate variability: A ratio Bland–Altman analysis on the log scale indicated a percent bias of 85.1% (App relative to H10) with ratio limits of agreement [−6.5%, 266.6%] (LoA 95% CIs: lower [−22.4%, 14.2%], upper [204.9%, 334.5%]). The proportional bias test on log-differences showed no evidence of proportional bias (slope = 0.041, $p = 0.787$). On the raw scale, MAE = 63.5 ms, MdAPE = 86.7%, and 8.3% of estimates were within ±20% of H10. CCC was 0.220.

Table 1: Summary statistics for Bland–Altman agreement (HR on the absolute scale; HRV on the ratio scale).

	n	Bias	LoA	MAE	RMSE	CCC
HR (bpm)	55	0.49	[−3.92, 4.90]	1.48	2.28	0.974
HRV (%)	48	85.1%	[−6.5%, 266.6%]	63.5 ms	–	0.220

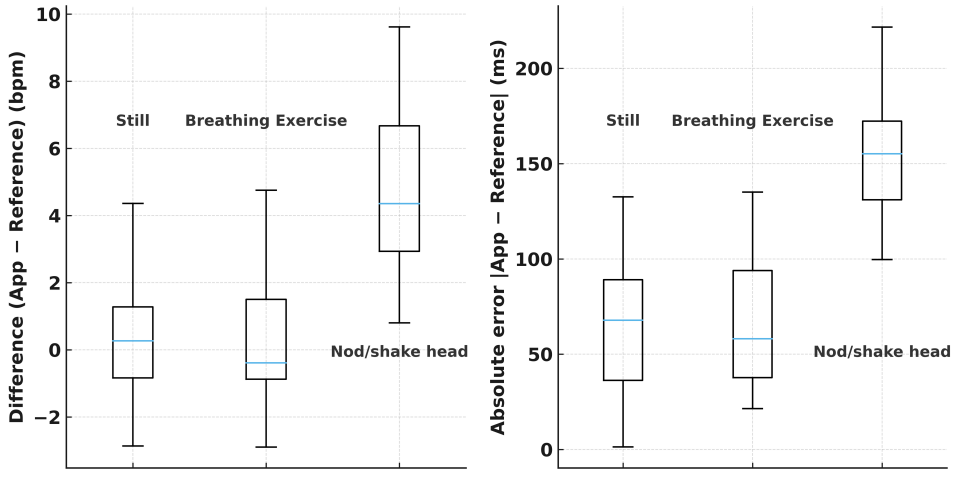

Figure 5: HR and HRV results by test type. Subjects were either still, conducting a breathing exercise, or asked yes or no questions to which they responded by nodding or shaking their heads.

Box-and-whisker plots (Figure 5) compare error distributions across recording types for HR (left; absolute difference) and HRV (right; absolute error in ms).

HR. Median (IQR) App–H10 difference by type: Still: 0.27 bpm [IQR -0.84 to 1.28], $n = 55$; Breathing Exercise: -0.39 bpm [IQR -0.87 to 1.50], $n = 10$; Nod/shake head: 4.35 bpm [IQR 2.93 to 6.67], $n = 10$. Kruskal–Wallis: $H = 18.14$, $p < 0.001$.

HRV. Median (IQR) absolute error by type: Still: 67.86 ms [IQR 36.30 to 89.06], $n = 48$; Breathing Exercise: 58.13 ms [IQR 37.75 to 93.92], $n = 8$; Nod/shake head: 155.21 ms [IQR 131.03 to 172.30], $n = 8$. Kruskal–Wallis: $H = 19.06$, $p < 0.001$.

Overall, HR errors remain small in still and breathing tasks but increase substantially during nod/shake movements. HRV absolute errors are largest in nod/shake head, consistent with motion-sensitivity of short-window RMSSD estimates.

3.2 Atrial Fibrillation Model

Table 2: Confusion matrix for the logistic AF classifier on the test set.

	Predicted Label	
True Label	Normal	AF
Actual Normal	108	12
Actual AF	4	156

The AF logistic regression model shows good performance with an overall accuracy of 94.2%. From table 2, it can be seen the model only mispredicted 4 patients with AF and 12 without, achieving a sensitivity of 97.5% and specificity of 89.6%. The F1 score is 95.1%.

4 Discussion

4.1 AF Results

Testing the AF classifying logistic regression model on the unseen patients yielded a 94.2% model accuracy. Previous works have developed CNNs to classify AF based on PPG data from the MIMIC PERform AF Dataset that was used to train and test the logistic regression model. The best CNN model for classifying AF on PPG data by Mäkynen et al. (2024) achieved an accuracy of 90.77%. From this, we can see that the feature-based logistic regression can outperform a deep learning model, despite its simplicity. Furthermore, the model was developed outside of the app environment. Using a simple model structure such as logistic regression with only 11 predictors allowed for an easy integration into the application, without the need for significant additional computational power. The confusion matrix in Table 2 gives us further insight into the model performance. Only 4 AF samples were misclassified, giving a sensitivity of 97.5%, meaning the model performs strongly at identifying AF. However, the specificity value is only 89.6%. These false positives are not ideal as they could lead to unnecessary alerts or follow-ups. On the other hand, the high recall for AF data is important and clinically valuable in that it will rarely miss an AF event. To understand where the model was producing errors, an investigation into the misclassified test set samples was conducted.

Figure 6 displays two of the misclassified samples. These time series clearly display poor signal quality. As these recordings were obtained from ICU patients using bedside monitors, some degree of noise and artefacts is expected. Since these short segments were extracted from longer 20-minute recordings, it is likely that certain periods contained degraded or unstable signals. Consequently, it is unreasonable to expect the model to perform accurately on data with such low fidelity, as these samples represent outliers that do not reflect the typical signal characteristics present in the training dataset or application. Removing these samples from the test data would increase the accuracy of the model and significantly improve the specificity.

Using a feature based model helped to overcome some of the challenges of using a PPG dataset to train a model that will be used on rPPG data. The MIMIC dataset contains PPG data recorded at a frequency of 125 Hz; however, a high-end iPhone only allows recording up to a maximum value of 120 Hz on the front-facing camera (Apple Inc., 2025a). The issue that arises here is the deep learning models are generally trained directly on the time series data to allow models to learn about important features. As a result, the data passed to the model must be recorded at the same frequency as the data on which the model was trained. The application output and dataset do not have matching frequencies; consequently, to implement an ML model into the app, the recorded data would have to be

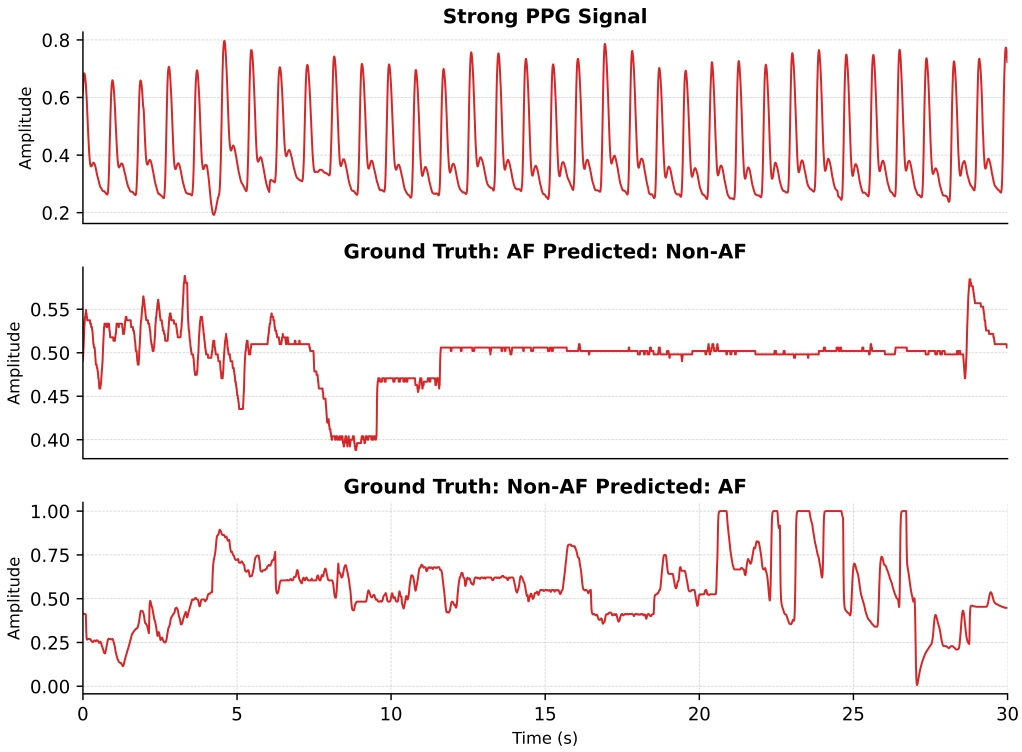

Figure 6: Example sample section from the MIMIC PERform AF dataset- one high-quality recording and two poor-quality cases misclassified by our model.

upscaled to 125 Hz. This functionality would have introduced unnecessary complexity to the system. Additionally, the app does not record at a consistent frame rate, meaning the data is not recorded at a consistent frequency. This aspect of the app would add an additional layer of complexity to up-scaling the rPPG signal after it is extracted from the video. Although this limitation in matching frequencies could be overcome by limiting the system to record at a lower frame rate, the issue of having to upscale data will always be present due to the limitations of using a high-frequency dataset. The logistic regression model overcomes these setbacks by using only features from the data that are independent of the frequency to classify a subject.

It is important to note that while the logistic classifier achieved a high level of accuracy, this figure does not reflect the performance of the complete application in classifying AF. Due to the absence of a labelled rPPG video database containing both AF and non-AF subjects, the end-to-end functionality of the system, from signal acquisition to AF classification, could not be fully validated. As access to patients with AF was not possible during this project, the model was trained and tested exclusively on PPG data. Although the model performs strongly on this dataset, with an accuracy of 94.2%, this result does not account for the additional sources of error introduced during rPPG signal acquisition, such as motion artefacts, lighting variability, and camera quality. This represents a key limitation of the project; however, its impact was mitigated by ensuring consistent data handling across both the mobile application and the PPG dataset used during model training. Furthermore, the application reports AF classification as a probabilistic confidence score rather than a binary output, enabling users to interpret the system's certainty in its predictions.

4.2 App Validation Discussion

The validation against the Polar H10 indicates that the application provides accurate HR estimates under appropriate conditions, while HRV remains sensitive to motion and short-window instability.

Heart rate On *Still* recordings, Bland–Altman analysis (Figure 4, Table 1) showed a near-zero mean difference of $\bar{d} = 0.49$ bpm with narrow 95% limits of agreement $[-3.92, 4.90]$ bpm, a high concordance (CCC = 0.974), and 98.2% of estimates within ±5 bpm of H10. These results support the app's suitability for routine HR monitoring in realistic resting conditions. However, error distributions differed by recording type (Figure 5): the median App–H10 difference was small during *Still* (0.27 bpm; IQR −0.84 to 1.28) and *Breathing Exercise* (−0.39 bpm; IQR −0.87 to 1.50), but increased during *Nod/shake head* (4.35 bpm; IQR 2.93 to 6.67) . This pattern is consistent with rPPG susceptibility to head motion (ROI shifts, skin-specular changes), which shows the importance of motion control and quality gating.

heart rate variability In contrast, HRV showed large dispersion. On *Still* recordings, ratio-scale Bland–Altman indicated a positive percent bias of 85.1% with very wide ratio limits of agreement [−6.5%, 266.6%] and low concordance (CCC = 0.220); only 8.3% of estimates fell within ±20% of H10. When summarised as absolute error on the by-condition box-plots, HRV was lowest for *Still* (median 67.86 ms; IQR 36.30–89.06), similar for *Breathing Exercise* (58.13 ms; IQR 37.75–93.92), and largest for *Nod/shake head* (155.21 ms; IQR 131.03–172.30), with a significant across-type difference ($H = 19.06$, $p < 0.001$). These findings reflect known challenges: beat-to-beat jitter from peak timing errors, frame-rate quantisation, and residual motion contaminate the IBI series, inflating RMSSD particularly in short windows.

The results of our validation show that HR is deployment-ready for resting use. However, difficulties remain under more challenging conditions, such as the head movements induced by the nodding/ shaking head tests. On the other hand, HRV requires caution in implementation, due to the large dispersion and low concordance limits. We suggest that HRV measurement quality may be improved with longer recording time, however, have not yet collected this data.

Limitations Although we limited participant guidance, all tests were conducted indoors under controlled conditions. Consequently, real-world end-to-end performance will depend on users following our measurement instructions and on environmental factors.

4.3 Further Research

While this paper successfully demonstrated the feasibility of real-time rPPG-based cardiovascular monitoring on a mobile platform, several limitations present opportunities for further investigation and improvement. Future work should focus on enhancing the robustness and clinical reliability of the system under real-world conditions. In particular, one of the most critical areas for further research is the accumulation of a comprehensive, labelled rPPG video dataset containing both AF and non-AF recordings.

Access to such a dataset would enable full end-to-end validation of the system, allowing the performance of the AF classifier to be evaluated on real rPPG signals rather than contact-based PPG data. This would provide a more accurate representation of how the application performs in practical use, especially when factors such as motion artefacts, lighting variations, and camera differences are present. Establishing such a dataset would likely require collaboration with clinical institutions to obtain video recordings from patients with confirmed AF diagnoses, ensuring appropriate ethical approval and patient consent. To increase the dataset's diversity and robustness, recordings should capture a range of lighting conditions, skin tones, camera types, and motion levels. Access to this type of comprehensive rPPG dataset would not only allow for more rigorous evaluation of AF detection performance but also support the development of more advanced machine learning models capable of handling the variability inherent in real-world mobile environments.

5 Conclusion

This study developed an iOS application that performs on-device, real-time extraction of rPPG signals to estimate heart rate, heart rate variability, and the probability of atrial fibrillation. The implementation demonstrates the feasibility of conducting end-to-end rPPG processing and lightweight classification on consumer mobile hardware, producing promising results under controlled conditions while maintaining low computational cost suitable for mobile deployment. The AF classifier achieved an accuracy of 94.2%, while the heart rate estimation yielded a mean absolute error of only 1.48 bpm, indicating strong performance in these two areas. However, the HRV estimation exhibited a larger mean absolute error of 63.5 ms, primarily due to noise and instability within the rPPG signal that, while not substantially affecting HR estimation, had a pronounced impact on HRV feature reliability.

However, the evaluation of the AF classifier was limited by the lack of a labelled rPPG dataset containing recordings of patients in atrial fibrillation. As a result, model performance was assessed using PPG data rather than end-to-end rPPG signals, which constrains the generalisability of the results. Future work should focus on the collection of a comprehensive, labelled rPPG video dataset that includes AF recordings across diverse subjects and conditions. In addition, improving signal quality would assist in reducing noise and artefacts, particularly improving HRV estimation. Addressing these limitations will help strengthen the reliability of the system across different use cases

AI Acknowledgement OpenAI's Codex product was used to assist in writing code and planning project architecture for the iOS app. In creating this paper, generative AI has been used to refine and edit written content, as well

as write code to generate figures.

References

Apple Inc. (2025a). iPhone 17 Pro and 17 Pro Max - technical specifications. https://www.apple.com/au/iphone-17-pro/specs/. Accessed 2025-10-19

Apple Inc. (2025b). Vision - Apple developer documentation. https://developer.apple.com/documentation/vision. Accessed 2025-10-16

Butterworth, S. (1930). On the theory of filter amplifiers. *Experimental Wireless and the Wireless Engineer* 7, 536–541

Charlton, P. H. (2022). Mimic perform datasets. doi:10.5281/zenodo.6807403

de Haan, G. and Jeanne, V. (2013). Robust Pulse Rate From Chrominance-Based rPPG. *IEEE Transactions on Biomedical Engineering* 60, 2878–2886. doi:10.1109/TBME.2013.2266196

Fitzpatrick, T. B. (1988). The validity and practicality of sun-reactive skin types i through vi. *Archives of Dermatology* 124, 869–871. doi:10.1001/archderm.124.6.869

Haugg, F., Elgendi, M., and Menon, C. (2022). Effectiveness of Remote PPG Construction Methods: A Preliminary Analysis. *Bioengineering (Basel)* 9, 485. doi:10.3390/bioengineering9100485

Huang, R.-Y. and Dung, L.-R. (2016). Measurement of heart rate variability using off-the-shelf smart phones. *BioMedical Engineering OnLine* 15, 11. doi:10.1186/s12938-016-0127-8

Li, K. H. C., White, F. A., Tipoe, T., Liu, T., Wong, M. C., Jesuthasan, A., et al. (2019). The current state of mobile phone apps for monitoring heart rate, heart rate variability, and atrial fibrillation: Narrative review. *JMIR mHealth and uHealth* 7, e11606. doi:10.2196/11606

Li, S., Elgendi, M., and Menon, C. (2024). Optimal facial regions for remote heart rate measurement during physical and cognitive activities. *npj Cardiovascular Health* 1, 33. doi:10.1038/s44325-024-00033-7

Linz, D., Gawalko, M., Betz, K., Hendriks, J. M., Lip, G. Y., Vinter, N., et al. (2024). Atrial fibrillation: epidemiology, screening and digital health. *The Lancet Regional Health – Europe* 37. doi:10.1016/j.lanepe.2023.100786

Liu, X., Fromm, J., Patel, S., and McDuff, D. (2020). Multi-task temporal shift attention networks for on-device contactless vitals measurement. *arXiv preprint arXiv:2006.03790*

Mäkynen, M., Ng, G. A., Li, X., Schlindwein, F. S., and Pearce, T. C. (2024). Compressed deep learning models for wearable atrial fibrillation detection through attention. *Sensors* 24. doi:10.3390/s24154787

Nesheiwat, Z., Goyal, A., and Jagtap, M. (2023). Atrial fibrillation. In *StatPearls* (Treasure Island (FL): StatPearls Publishing). [Updated 2023 Apr 26]

Niu, X., Han, H., Shan, S., and Chen, X. (2018). VIPL-HR: A multi-modal database for pulse estimation from less-constrained face video. In *Asian Conference on Computer Vision* (Springer), 562–576

Odinaev, I., Wong, K. L., Chin, J. W., Goyal, R., Chan, T. T., and So, R. H. Y. (2023). Robust heart rate variability measurement from facial videos. *Bioengineering (Basel)* 10, 851. doi:10.3390/bioengineering10070851

Schaffarczyk, M., Rogers, B., Reer, R., and Gronwald, T. (2022). Validity of the Polar H10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women. *Sensors* 22, 6536. doi:10.3390/s22176536

Sun, Z., Junttila, J., Tulppo, M., Seppanen, T., and Li, X. (2022). Non-contact atrial fibrillation detection from face videos by learning systolic peaks. *IEEE Journal of Biomedical and Health Informatics* 26, 4587–4598. doi:10.1109/JBHI.2022.3193117. Epub 2022 Sep 9

Wang, W., den Brinker, A. C., Stuijk, S., and de Haan, G. (2017). Algorithmic Principles of Remote PPG. *IEEE Transactions on Biomedical Engineering* 64, 1479–1491. doi:10.1109/TBME.2016.2609282