CardioVision: A Multi-Modal AI System for Real-Time Cardiac Arrest Risk Prediction

Authors: Lukhsaan Elankumaran, Varun Gande, Nirjhar Dey
Mentors: Jaemar Miller, Vineel Nagisetty
RBC Borealis Let's Solve It! Spring 2025 Cohort

1. Executive Summary

Sudden cardiac arrest (SCA) is a life-threatening condition that often occurs without warning (Myerburg & Castellanos, 2003), outside of clinical settings where timely intervention is possible. Current detection methods are largely reactive (Zipes & Wellens, 1998), relying on post-event ECG interpretation or in-hospital monitoring, leading to critical delays in care. CardioVision addresses this challenge with a real-time, AI-powered system that predicts cardiac arrest risk using physiological data from wearable devices like the Apple Watch.

Developed during the Spring 2025 RBC Borealis Let's Solve It! program, CardioVision integrates multiple signal types – including Electrocardiogram (ECG) waveforms and HealthKit-derived metrics like Heart Rate (HR), Resting Heart Rate (RHR), Heart Rate Variability (HRV), and High Heart Rate Events (HHR) (Apple Inc., 2023) – to deliver dynamic cardiac risk scores (Low, Medium, or High) in real time.

A Random Forest model serves as the initial decision layer, classifying health risk based on the four key physiological metrics (HR, HRV, RHR, and HHR). This model provides a lightweight, high-speed preliminary filter to determine whether the system should prompt the user to collect an ECG for further analysis. It allows for continuous background monitoring without overwhelming the user or backend with unnecessary ECG requests, preserving energy and enhancing usability.

The system's core is a fine-tuned Bidirectional LSTM (BiLSTM) that classifies risk based on beat-segmented ECG waveforms. Supporting submodels include:

- A Random Forest Model for the Initial Metric Classification
- A Rule-Based Threshold model for HR
- A Logistic Regression classifier for RHR
- A Logistic Regression Meta-Classifier for an ensemble of XGBoost, CatBoost, and LightGBM models for HRV
- A Random Forest for HHR classification
- A Bi-Directional LSTM for ECG waveforms

These models feed into a meta-learner BiLSTM ensemble that aggregates outputs for final risk prediction.

CardioVision was trained and validated on clinical datasets including MIT-BIH Arrhythmia, INCART, Holter, and MIMIC-III waveform data (Goldberger et al., 2000; Moody et al., 2001; Johnson et al., 2016), alongside simulated HealthKit metrics based on medical guidelines. The OHCA dataset (Out-of-Hospital Cardiac Arrest) was reserved strictly for testing to evaluate real-world

performance. Model training employed Focal Loss, SMOTE, and learning rate scheduling to address class imbalance and ensure generalization. ECG data was augmented via noise, scaling, and shifting, and final models were quantized for edge deployment.

The resulting system achieved 97.72% classification accuracy, with high recall on high-risk patients. A real-time backend using FastAPI, and a fully functional watchOS/iOS prototype built with Swift and HealthKit, demonstrate CardioVision's readiness for deployment in wearable and clinical environments.

CardioVision exemplifies how ensemble AI models and wearable health data can enable proactive cardiac monitoring. Its real-time performance, modular architecture, and accuracy across diverse datasets mark it as a strong candidate for integration into next-generation digital health ecosystems.

2. Introduction and Background

2.1. The Challenge of Predicting Cardiac Arrest

Despite the growing accessibility of health monitoring technologies, early prediction of cardiac arrest remains an unresolved challenge. While conditions like hypertension or arrhythmias can be flagged through regular screening, cardiac arrest is often preceded by subtle or undetected physiological changes (Zipes & Wellens, 1998; Myerburg & Castellanos, 2003). Existing solutions tend to react after symptoms appear – missing the opportunity to alert users before an event becomes critical. A true preventative system must be able to interpret multiple real-time physiological signals and translate them into actionable insights for both patients and healthcare providers.

2.2. The Rise of Wearables in Preventive Cardiology

Consumer health devices like the Apple Watch and Fitbit have introduced new possibilities (Apple Inc., 2023; Hernandez-Silveira et al., 2015) in continuous cardiac monitoring. These wearables collect physiological metrics such as heart rate (HR), resting heart rate (RHR), heart rate variability (HRV), and single-lead ECG signals. However, most wearable applications focus on fitness tracking or basic arrhythmia detection (Perez et al., 2019) and lack sophisticated systems for predicting cardiac events before symptoms manifest. CardioVision leverages the ubiquity and real-time data accessibility of these devices to close this gap in preventative cardiac care.

2.3. Data-Driven Approach to Real-Time Risk Prediction

To develop a robust and generalizable risk prediction system, CardioVision was trained and evaluated on a diverse collection of clinical-grade datasets. These include:

- MIT-BIH Arrhythmia Database – annotated half-hour ECG recordings with multi-class arrhythmia labels
    - Referred to as MIT-BIH
- St. Petersburg INCART 12-Lead Arrhythmia Database – multilead ECG signals from a variety of arrhythmias
    - Referred to as INCART
- Sudden Cardiac Death Holter Database – long-duration ECGs of patients with sudden cardiac death
    - Referred to as Holter
- MIMIC-III Waveform Database – ICU physiological signals including ECG and vital signs
    - Referred to as MIMIC-III
- Out-of-Hospital Cardiac Arrest (OHCA) Dataset – used to simulate real-world cardiac arrest scenarios
    - Referred to as OHCA

In parallel, simulated Apple HealthKit data was generated to reflect physiological metrics expected from wearables, based on clinical ranges provided by health institutions.

This data was used to train specialized machine learning models that process different physiological signals and collaboratively determine a user's cardiac arrest risk in real time.

2.4. Project Scope and Deployment Goals

CardioVision was built over an intensive 8-week development cycle as part of the RBC Borealis Let's Solve It! Spring 2025 program. The project's goal was to create a deployable, real-time cardiac risk prediction system that could operate on both mobile and backend infrastructure. The final prototype includes a fine-tuned AI model, a FastAPI-based backend, and a functional watchOS/iOS frontend built in Swift and integrated with Apple HealthKit. The system simulates real-time predictions by continuously monitoring vital signs and, when needed, prompting users to record an ECG – enabling preventative care through timely alerts and risk assessments.

CardioVision represents a scalable step forward in wearable cardiac health technology – uniting clinical insight, real-time signal processing, and edge-device readiness in one cohesive system.

3. Description of the Problem

3.1. Core Problem: Delayed Detection of Cardiac Arrest

Cardiac arrest frequently strikes without warning. Research shows that out-of-hospital cardiac arrest (OHCA) affects approximately 4.4 million individuals each year globally, with survival rates below 10% for unwitnessed events (Berdowski et al., 2010; Grasner et al., 2020). Existing detection systems – such as isolated ECG readings or hospital telemetry – are reactive by nature (Zipes & Wellens, 1998) and unable to capture early warning patterns across multiple physiological signals.

3.2. Causes: Fragmented Data & Lack of Real-Time Context

One major limitation in current predictive systems is the overreliance on isolated features. Many AI models, particularly convolutional neural networks (CNNs) or LSTMs trained solely on ECG signals, often demonstrate limited predictive value, with area under the curve (AUC) scores frequently falling below 0.6 when analyzing individual metrics (van de Leur et al., 2021; Faust et al., 2018). Compounding this issue is the fragmented nature of clinical data, where vital physiological indicators like ECG, heart rate (HR), and heart rate variability (HRV) are typically stored and analyzed in isolation – hindering the ability to contextualize temporal trends and interactions across modalities (Shickel et al., 2018). Additionally, while wearables collect HR, HRV, and ECG data, this is rarely done in a coordinated or continuous manner that reflects dynamic health states. HRV, in particular, has shown only modest predictive power when used on its own to assess cardiac risk, limiting its effectiveness without multimodal integration (Shaffer & Ginsberg, 2017).

3.3. Impact: High Mortality and Limited Preventative Care

Cardiac arrest remains a major global health challenge, claiming hundreds of thousands of lives annually. Survival is heavily dependent on rapid detection and immediate intervention, with even brief delays significantly lowering the chance of recovery (Benjamin et al., 2019; Myerburg & Castellanos, 2003). Despite

advances in health technology, there is currently no widely adopted system to proactively alert individuals at risk before a cardiac arrest occurs. As a result, over 70% to 90% of victims die before ever reaching hospital care (Grasner et al., 2020). This lack of early warning not only limits patient outcomes but also places immense strain on emergency response systems, reducing opportunities for preemptive treatment and increasing the burden on already stretched clinical resources.

3.4. Gaps in Current Models: Lack of Multimodal, Temporal Integration

Recent studies have demonstrated that multimodal models that combine ECG with additional physiological signals, such as EEG, significantly outperform single-signal systems. These integrated approaches often achieve area under the curve (AUC) values between 0.8 and 1.0, while models relying solely on ECG typically perform below an AUC of 0.6 (Xu et al., 2022; Wang et al., 2021). Even in intensive care settings, where classical HRV-based models powered by algorithms like LightGBM have achieved respectable AUCs of around 0.88 (Chen et al., 2022), these systems often fail to capture the dynamic interplay between heart rate and ECG signals over time. CardioVision addresses this limitation by integrating sequential ECG inputs – processed through a Bidirectional LSTM – with machine learning models for HR, HRV, RHR, and HHR. These submodels are unified within a meta-learner ensemble that enables proactive, real-time, and multimodal cardiac risk assessment. This approach moves beyond the reactive nature of existing models, aiming instead to deliver predictive alerts before a critical event occurs.

4. Criteria for Acceptable Solutions

To be viable in real-world, non-clinical environments, any cardiac arrest prediction system must meet a specific set of technical and clinical criteria. These criteria address the limitations of existing solutions and enable safe, accurate, and deployable decision-making.

4.1. Real-Time Compatibility

The system must support low-latency processing to respond to rapidly changing physiological states. Even minor delays in cardiac event detection can reduce the effectiveness of intervention (Johnson et al., 2016). Real-time compatibility requires asynchronous data ingestion, immediate risk classification, and minimal overhead during inference – especially for wearable or streaming applications.

4.2. High Recall for High-Risk Cases

In predictive healthcare, especially for life-threatening events like cardiac arrest, sensitivity (recall) is paramount. False negatives can have catastrophic consequences, so modern predictive models often prioritize recall over specificity. Techniques such as Focal Loss (Lin et al., 2017) and class weighting are frequently employed to reduce bias toward the majority class and emphasize correct classification of high-risk cases.

4.3. Multi-Modal Input Support

Cardiac arrest is not identifiable from a single biomarker. Integrating diverse physiological signals – including ECG, HR, HRV, RHR, and high heart rate events – allows a model to detect subtle but critical trends. Research on multimodal systems has demonstrated AUC improvements of 20–30% when combining sequential physiological inputs versus using

ECG alone (Liu et al., 2021). Temporal dependencies across modalities are especially valuable when tracked continuously.

4.4. Portability and Edge Compatibility

Deployment in consumer environments demands efficient models that can run on mobile or embedded devices. CardioVision's ECG BiLSTM was optimized via static quantization, which reduces model size and accelerates inference while preserving accuracy. Quantized deep learning models have been shown to achieve up to 4× compression and 2× speedup with minimal accuracy loss (Jacob et al., 2018), making them suitable for edge inference on smartwatches or mobile apps.

Meeting these criteria ensures a cardiac risk prediction system can function reliably in real-time, adapt to diverse data types, prioritize patient safety, and be deployed at scale in wearable and clinical settings.

5. Proposed Solution: CardioVision

CardioVision is a real-time, machine learning-based system designed to predict cardiac arrest risk using both clinical ECG data and wearable physiological signals. The solution integrates a modular ensemble of specialized models, each responsible for analyzing a different biomarker associated with cardiac health. By combining these outputs through an adaptive meta-learner to fine-tune the BiLSTM, CardioVision provides accurate, real-time predictions that categorize risk into Low, Medium, or High – enabling early intervention before critical cardiac events occur.

The solution was developed to meet key clinical and technical requirements: high recall for high-risk individuals, support for multimodal inputs (HR, HRV, RHR, HHR, ECG), real-time inference, and compatibility with mobile/embedded systems such as the Apple Watch. The architecture, training strategies, and deployment goals were designed with these constraints in mind.

5.1. Model Architecture

CardioVision's architecture follows a dual-stage pipeline designed to mimic clinical triage: an initial risk classification layer for quick screening, followed by a fine-tuned ensemble model for high-fidelity prediction. This modular approach enables real-time responsiveness while preserving interpretability and robustness across diverse cardiac metrics..

5.1.1. Initial Risk Model

The first stage of CardioVision's architecture is a lightweight Random Forest classifier that evaluates incoming physiological metrics from wearables. These include:

- Heart Rate (HR)
- Heart Rate Variability (HRV)
- Resting Heart Rate (RHR)
- High Heart Rate Events (HHR)

Trained on simulated Apple HealthKit data, this model categorizes users into *risk* and *no-risk* groups in real time. If classified as *at risk*, the system prompts the user to collect an ECG. This initial triage conserves device energy and avoids unnecessary user interventions while maintaining high sensitivity to early warning signals.

5.1.2. Final Risk Model

If an ECG is collected, CardioVision advances to its second stage: a five-submodel ensemble that analyzes physiological and signal-based features to assign a final risk label (*Low*, *Medium*, or *High*). Outputs from each submodel are consolidated by a fine-tuned Bidirectional LSTM (BiLSTM) trained using Focal Loss, designed to reduce false negatives – especially in high-risk cases.

Submodel Breakdown:

- Electrocardiogram (ECG) – BiLSTM Network

A 3-layer Bidirectional LSTM processes beat-segmented ECG waveforms to detect temporal risk patterns. This deep neural network serves as the backbone of the prediction pipeline, classifying input sequences into one of three risk levels.

- Heart Rate (HR) – Rule-Based Threshold Model

A rule-based model flags heart rate values outside clinically defined bounds (e.g., <40 bpm or >110 bpm). This non-ML model offers fast alerts for extreme HR deviations.

- Resting Heart Rate (RHR) – Logistic Regression

This model captures deviations from baseline resting rates to flag cardiovascular irregularities. It is useful for identifying abnormal variations during periods of rest.

- Heart Rate Variability (HRV) Gradient Boosting Ensemble

A voting ensemble combining XGBoost, LightGBM, and CatBoost, trained on RR interval-derived features (RMSSD, SDNN, pNN50, etc.). It reflects autonomic nervous system function and stress patterns.

- High Heart Rate Events (HHR) – Random Forest Classifier

This model identifies sustained, abnormal HR spikes that may indicate arrhythmic events. It factors in spike count, duration, and deviation magnitude.

5.1.3. Fine-tuned BiLSTM (Final Predictor)

The output vectors from the five submodels – alongside beat-segmented ECG data – are passed to a fine-tuned BiLSTM trained with Focal Loss ($\alpha = 0.95$, $\gamma = 2.0$) and class weighting. This approach penalizes misclassification of minority classes (especially high-risk) and improves recall. The fine-tuning dataset was curated using feedback samples (true positives and false negatives) from MIT-BIH, Holter, MIMIC-III, and INCART databases, augmented with SMOTE to ensure class balance. This ensures the final risk classifier is both accurate and highly sensitive to life-threatening conditions.

5.1.4. Real-Time Inference Pipeline

CardioVision's real-time pipeline is designed for continuous monitoring and decision-making across wearable and backend systems. The complete flow is as follows:

The watchOS frontend continuously collects HealthKit metrics – including Heart Rate (HR), Resting Heart Rate (RHR), Heart Rate Variability (HRV), and High Heart Rate Events (HHR) – directly from the user's Apple Watch.

These metrics are sent to the FastAPI backend as a JSON payload, where they are processed by an initial Random Forest model trained to identify potential cardiac risk using only physiological (non-ECG) inputs.

The backend returns an initial risk score (No Risk or Risk) to the frontend.

- If No Risk is detected, the monitoring cycle continues silently in the background.

- If Risk is detected, the frontend prompts the user to record an ECG using the Apple Watch's HealthKit ECG interface.

Once the ECG is recorded, the signal is segmented and sent to the backend via another JSON request. The backend then:

- Passes the ECG to a fine-tuned BiLSTM model, which has been trained with Focal Loss to reduce false negatives.

- References the outputs of the submodels for HR, HRV, RHR, and HHR to provide contextual understanding of the event to reduce false negatives.

The BiLSTM model produces a final cardiac arrest risk classification – Low, Medium, or High – which is sent back to the frontend and displayed to the user in real time.

This end-to-end architecture ensures rapid, responsive risk assessment with minimal delay, integrating wearable sensing, backend intelligence, and real-time alerts into a unified pipeline.

5.2. Data Sources

The CardioVision system was developed using a diverse collection of clinical and simulated datasets, selected to ensure both physiological accuracy and real-world applicability. Clinical ECG datasets were sourced from leading open-access repositories (Goldberger et al., 2000; Moody et al., 2001; Johnson et al., 2016; Lagani et al., 2015) and reflect a wide range of patient demographics, signal durations, lead configurations, and cardiac conditions. These were used for training, validation, and generalization testing across various submodels. Simulated wearable data was also incorporated to replicate the signal patterns (American Heart Association, 2023) and formats encountered in consumer health devices, enabling real-time testing and frontend integration. The combination of long-form Holter data, ICU waveforms, high-resolution multilead recordings (Lagani et al., 2015; Johnson et al., 2016; Moody et al., 2001), cardiac arrest events, and structured HealthKit-style inputs provided a comprehensive foundation for model development. Each dataset is described in detail below, including its structure, sampling rate, annotation schema, and relevance to the project's goals.

5.2.1. Apple HealthKit & Mock HealthKit Data (Wearable Simulation)

As wearable technology becomes increasingly central to preventative health monitoring, CardioVision incorporates simulated Apple HealthKit data to reflect the physiological inputs typically collected from consumer-grade devices such as the Apple Watch (Apple Inc., 2023). This component was essential for building and validating a system capable of ingesting real-time signals, simulating deployment in a non-clinical setting.

The Apple HealthKit module provides critical real-time inputs to CardioVision's pipeline. These include:

- Heart Rate (HR)
- Resting Heart Rate (RHR)
- Heart Rate Variability (HRV)
- High Heart Rate Events (HHR)

These metrics are sent to the FastAPI backend in JSON format, enabling seamless integration with the frontend (watchOS app) and simulating live data transmission from wearable sensors. By testing the full input-output flow without requiring physical hardware, the system remains testable and repeatable in development environments.

5.2.1.1. Data Generation Approach

To simulate HealthKit data, a Python script was developed to generate mock physiological samples, categorized into “risk” and “no risk” datasets. Each category contains 100 JSON files, with values drawn from clinically validated ranges:

- Resting Heart Rate (RHR):

Normal values fall between 55–85 beats per minute for adults, with elevated rates (>90 bpm) often associated with increased cardiac risk or poor cardiovascular conditioning. These ranges were derived from sources such as the American Heart Association, Harvard Health Publishing, and WebMD (American Heart Association, 2023; Harvard Health Publishing, 2023; WebMD, 2023).

- Heart Rate Variability (HRV):

HRV values above 50 ms (typically measured via RMSSD or SDNN) are considered healthy, while readings below 40 ms indicate stress or sympathetic dominance – a known risk factor for sudden cardiac events. These ranges were based on standards from Kubios HRV and recent cardiovascular studies.

- Heart Rate (HR):

Typical resting HR ranges between 60–100 bpm. Simulated elevated heart rate values (110–140 bpm) (American Heart Association, 2023) represent sustained tachycardic events, especially when not correlated with physical activity.

- High Heart Rate Events (HHR):

This metric tracks the number of distinct high-rate episodes. In risk samples, 1–5 events are recorded; in no-risk samples, none occur.

Each generated file contains these metrics in the following structure:

```
{
    "rhr": 93.5,
    "hrv": 27.1,
    "hr": 131.4,
    "hhr": 2
}
```

This format mirrors real-time JSON payloads that might be retrieved from HealthKit in a production Apple Watch app, ensuring consistency across pipeline stages.

5.2.1.2. System Integration and Testing

The mock HealthKit dataset plays a pivotal role in validating the entire CardioVision pipeline. It allows for rigorous testing of the FastAPI backend’s real-time ingestion capabilities, ensuring that physiological inputs are parsed and routed correctly. Additionally, it enables verification of the threshold-based and machine learning submodels responsible for analyzing HR, HRV, RHR, and HHR. On the frontend, the dataset supports evaluation of the user interaction flow – including ECG recording prompts and risk status updates. By clearly distinguishing between “risk” and “no risk” cases, the mock data also facilitates robust testing of model recall, precision, and alert thresholds in a controlled yet clinically meaningful context.

5.2.1.3. Clinical Grounding and Risk-Exercise Differentiation

By anchoring the simulated values in ranges drawn from clinical literature (American Heart Association, 2023; Kubios Oy, 2022; Harvard Health Publishing, 2023; WebMD, 2023), the system ensures that simulated inputs reflect meaningful distinctions between at-risk and healthy individuals. This ensures that submodels trained or tested using mock data behave similarly when deployed in real-world use cases. To further distinguish cardiac risk from exercise-induced heart rate elevation, the system considers patterns across HR, HRV, RHR, and HHR. For instance, high HR accompanied by preserved HRV and low RHR typically indicates exercise, whereas elevated HR with suppressed HRV and persistently high RHR signals potential cardiac distress.

5.2.2. MIT-BIH Arrhythmia Database

The MIT-BIH Arrhythmia Database is a foundational clinical resource (Goldberger et al., 2000; Moody & Mark, 2001) widely used in ECG analysis and arrhythmia detection research. Maintained by MIT and the Beth Israel Hospital, it provides high-quality, annotated ECG data essential for training and evaluating signal-based cardiac models.

5.2.2.1. Database Overview

The dataset comprises 48 half-hour ambulatory ECG recordings drawn from 47 subjects, capturing diverse cardiac rhythms with two-channel signal recordings at a sampling rate of 360 Hz and 11-bit resolution over a ±10 mV range. These signals include common arrangements like Lead II and Lead V5, offering both rhythm and morphological information crucial for downstream analysis.

Approximately 110,000 beat annotations were manually generated by expert cardiologists (Moody & Mark, 2001), who identified each beat as normal, arrhythmic, or artifact. This rich annotation enables precise supervised training for models detecting anomalous cardiac events.

5.2.2.2. File Structure

The database adheres to the WFDB format, containing the following files for each recording:

- .dat: 2-channel raw ECG waveform at 360 Hz
- .hea: Header metadata including sampling rate, signal labels, and gain
- .atr: Beat-level annotations in line with AAMI standards
- .xws: Extended waveform markers for advanced waveform analysis

This structure supports robust signal preprocessing, beat segmentation, and feature extraction workflows.

5.2.2.3. Justification and References

MIT-BIH is recognized as a benchmark database (Goldberger et al., 2000) due to its clinical diversity, annotation rigor, and widespread usage in arrhythmia detection benchmarks. It was the first freely available standard test dataset designed for evaluating arrhythmia detectors and remains a reliable source for ECG-based cardiac critical event diagnosis.

5.2.3. MIMIC-III Waveform Database

The MIMIC-III Waveform Database is a large-scale, open-access repository of physiological signals collected in real-time from ICU patients. Developed through a collaboration between the MIT Laboratory for Computational Physiology and the Beth Israel Deaconess Medical Center (Johnson et al., 2016), it is one of the most comprehensive clinical signal datasets available for public use, widely adopted for machine learning and critical care research.

5.2.3.1. Database Overview

This dataset includes over 67,000 waveform recordings from approximately 30,000 adult ICU patients. A matched subset of over 22,000 records is aligned with detailed clinical data (Johnson et al., 2016), allowing temporal correlation between physiological signals and hospital events such as lab tests, medications, and procedures (Johnson et al., 2016). Signals are continuously sampled at 125 Hz and include electrocardiogram (ECG), arterial blood pressure (ABP), respiration, photoplethysmogram (PPG), and oxygen saturation (SpO_2). These recordings span several hours to multiple days per patient, providing high-resolution and temporally rich data reflective of dynamic ICU environments.

5.2.3.2. File Structure

All recordings are formatted using the standard WFDB (WaveForm DataBase) structure:

- .dat: Raw multi-channel waveform signals, often including ECG, ABP, and SpO_2
- .hea: Header files containing metadata such as sampling rate, signal names, gain factors, and number of samples
- n.dat / n.hea: Supplementary numeric files capturing discretely sampled vital signs, including heart rate and oxygen saturation, at regular time intervals

This structure facilitates consistent parsing, synchronization, and preprocessing of signals for downstream analysis.

5.2.3.3. Justification and References

MIMIC-III is widely regarded as a gold-standard dataset for modeling real-world clinical conditions. Unlike curated benchmark ECG datasets, it includes inherent noise and signal variability due to patient movement, sensor artifacts, and clinical interventions. This makes it an ideal training resource for models expected to operate under hospital-grade complexity and unreliability. Its linkage to rich clinical metadata further allows researchers to study signal-to-event relationships in real-time, elevating its value for predictive modeling in healthcare contexts (Johnson et al., 2016; Moody et al., 2020).

5.2.4. Sudden Cardiac Death Holter Database

The Sudden Cardiac Death Holter Database (SCDH) is a specialized collection of long-term ECG recordings curated to study patients at risk of sudden cardiac death (SCD). Developed by the Boston Beth Israel Hospital and distributed through PhysioNet (Goldberger et al., 2000), it provides detailed, continuous cardiac monitoring data under real-world ambulatory conditions.

5.2.4.1. Database Overview

This dataset consists of 23 full-day (24-hour) Holter ECG recordings from patients identified as being at elevated risk of SCD. The recordings are sampled at 250 Hz and capture long-duration cardiac behavior in a non-clinical, ambulatory setting. These records include cases from both survivors and non-survivors, offering a spectrum of cardiac rhythms ranging from normal sinus rhythm to life-threatening arrhythmias.

The recordings feature high temporal resolution and natural variability, including circadian rhythms, postural effects, and spontaneous arrhythmias, making the dataset especially suitable for training models to recognize prolonged cardiac stress and subtle anomalies over extended timescales.

5.2.4.2. File Structure

Data is stored using the WFDB format:

- .dat: Raw ECG waveform data sampled at 250 Hz over a 24-hour period
- .hea: Header files specifying the sampling rate, channel names, gain, and number of samples
- .atr: Annotation files labeling each beat as normal, arrhythmic, or artifact, using standard AAMI beat codes

This structure allows for efficient signal processing and segmentation for tasks such as beat classification, rhythm trend analysis, and long-term risk profiling.

5.2.4.3. Justification and References

The Sudden Cardiac Death Holter Database is particularly valuable due to its 24-hour continuous format, which mirrors real-world monitoring scenarios more closely than short-form ECG datasets. It enables detection of episodic, late-onset, or cumulative anomalies that may not appear in short clinical snapshots. Moreover, its focus on patients with known SCD risk ensures that the data captures pathophysiological signals associated with extreme cardiac outcomes, providing a rare and critical training resource for machine learning models targeting life-threatening conditions (Goldberger et al., 2000).

5.2.5. St. Petersburg INCART 12-Lead Arrhythmia Database

The St. Petersburg INCART 12-Lead Arrhythmia Database is a clinically annotated ECG dataset containing multilead recordings designed to capture a wide variety of arrhythmias under real-world diagnostic conditions. Developed by the Institute of Cardiological Technics (INCART) in Russia and distributed via PhysioNet (Goldberger et al., 2000), the dataset is widely recognized for its inclusion of full 12-lead ECG configurations and extensive beat-level annotation, making it uniquely suited for generalization testing in ECG-based machine learning systems.

5.2.5.1. Database Overview

This dataset comprises 75 annotated ECG recordings, each 30 minutes in duration, collected from 32 patients. All signals are sampled at 257 Hz and recorded simultaneously across 12 standard clinical leads (I, II, III, aVR, aVL, aVF, V1–V6). Together, these recordings contain more than 175,000 manually labeled beats, encompassing a broad range of cardiac pathologies.

Patient conditions represented in the dataset include atrial fibrillation, supraventricular tachycardia (SVT), myocardial infarction (MI), bundle branch blocks, ischemic episodes, and various forms of ventricular and supraventricular ectopy. The presence of multilead data enhances diagnostic richness, enabling morphology analysis across spatially distributed perspectives of the heart's electrical activity.

5.2.5.2. File Structure

Data is provided in WFDB format, consistent with standard clinical signal repositories:

- .dat: Raw 12-channel ECG waveform data sampled at 257 Hz
- .hea: Header metadata including channel labels, gain values, sampling frequency, and patient demographics
- .atr: Annotation files indicating beat locations and rhythm classifications, using AAMI-standard codes for arrhythmias and artifacts

This structure allows for synchronized multilead signal processing and supports a variety of downstream tasks, including

lead-wise feature extraction, rhythm classification, and spatial pattern learning.

5.2.5.3. Justification and References

The INCART database is particularly valuable for its comprehensive multi lead configuration, which is absent in many benchmark ECG datasets. The inclusion of simultaneous 12-lead data supports model generalization across leads and conditions, and facilitates validation of systems intended for hospital-grade ECG input. The dataset's emphasis on clinically diverse arrhythmias also ensures that models trained or tested on this resource are exposed to a broad range of physiologically distinct cardiac events. Its high annotation quality and real-world signal conditions make it an essential benchmark for arrhythmia classification research (Goldberger et al., 2000).

5.2.6. Out-of-Hospital Cardiac (OHCA) Database

The Out-of-Hospital Cardiac Arrest (OHCA) Database is a specialized clinical dataset designed to support the study of cardiac arrest events occurring outside traditional healthcare settings. Collected from emergency medical services (EMS) interventions (Acharya et al., 2018; Goldberger et al., 2000), this dataset captures the critical moments before and after defibrillation and is uniquely suited for evaluating high-risk prediction models under real-world emergency conditions.

5.2.6.1. Database Overview

The dataset consists of 260 anonymized ECG recordings collected from patients who experienced out-of-hospital cardiac arrest and received defibrillation treatment by EMS responders. Each record includes a 9-second segment of pre-shock ECG and a 1-minute segment of post-shock ECG. For our analysis, we specifically focused on the 9-second pre-shock segment, as it captures the heart's electrical activity during the critical moments of cardiac arrest, providing vital data for detecting high-risk patterns in real time.

Each case is labeled by a panel of three board-certified cardiologists (Acharya et al., 2018) according to its clinical outcome. Labels fall into two categories: ROEA (Return of Effective Activity), indicating a successful defibrillation and return of cardiac function, or NoROEA, indicating that defibrillation failed to restore a viable rhythm. These outcome labels allow the dataset to support binary classification tasks tied directly to life-or-death treatment response scenarios.

5.2.6.2. File Structure

The OHCA dataset is available in a mixed file format structure:

- .pdf: Scanned printouts of ECG waveform images. These files are anonymized, printed from original EMS equipment, and then scanned to preserve waveform shape and visual context.
- .txt: Digitized waveform data extracted from the ECG scans. These files contain amplitude-time series suitable for computational analysis and machine learning workflows.
- .xls: Excel spreadsheets containing pre-extracted features, including time-domain metrics (e.g., heart rate), frequency-domain parameters (e.g., spectral energy), wavelet coefficients, and nonlinear features (e.g., entropy, fractal dimension). These structured values allow rapid model input without requiring raw signal preprocessing.

5.2.6.3. Data Formatting for Real-Time Simulation

To support real-time demonstrations and compatibility with CardioVision's HealthKit-based frontend, a preprocessing script was developed to convert OHCA ECG recordings into a structure that mimics the Apple HealthKit ECG format. The original OHCA .txt waveform files, sampled at 250 Hz, were resampled to 512 Hz to match the frequency expectations of HealthKit-compatible processing pipelines.

The script systematically reads ECG text files from categorized subfolders (ROEA, noROEA, indeterminable), rescales the signals, and packages each one into a standardized JSON format that includes metadata such as startTime, samplingFrequency, and a list of resampled voltage values. These JSON files are saved and used to simulate high-risk ECG input during live system demos. An example file would be as follows:


```
{
        "startTime":
        "2025-05-04T10:03:47Z",

        "samplingFrequency": 512,

        "voltages": [0.1381,
        0.12551771016897476,
        0.05947606783740376,
        0.008112119939711802,
        -0.13320670402789073,
        -0.11459054421716428,
        -0.10731667693981818,
        -0.10661865322603141,
        -0.09439338773387089,
        -0.0754021300287101,
        ……,
        ,-0.02864062423408863,
        -0.0024434083676810603,
        0.07539010486376356]

}
```


This approach ensures that OHCA waveform data – originally formatted for clinical review – can be interpreted by CardioVision's mobile interface in real time, offering a realistic simulation of emergency cardiac events in wearable use cases. The conversion preserves the morphological integrity of the original ECG signals while enabling compatibility with the pipeline's frontend infrastructure.

5.2.6.4. Justification and References

The OHCA dataset holds significant clinical relevance due to its focus on cardiac arrest events in uncontrolled, prehospital environments. Unlike datasets collected under stable monitoring conditions, these recordings capture signal characteristics amid physical stress, motion, and rapid emergency response (Acharya et al., 2018; Goldberger et al., 2000). Furthermore, the inclusion of outcome labels allows for the evaluation of model performance in classifying life-threatening states, such as distinguishing between recoverable and non-recoverable rhythms. The dataset's pairing of raw waveform, expert annotation, and derived features provides a multifaceted input for model testing and external generalization validation (Acharya et al., 2018; Goldberger et al., 2000).

5.3. Training & Evaluation

Each submodel within the CardioVision system was trained and evaluated independently before being integrated into the final ensemble. This modular approach allowed for tailored preprocessing, model selection, and optimization techniques specific to the characteristics of each physiological signal. The training phase involved carefully curated datasets, robust validation strategies, and domain-informed augmentation pipelines to ensure generalizability across both clinical and wearable contexts. PyTorch was used as the primary framework for implementing and

training various models for waveform classification, enabling fine-grained control over the architecture, optimization process, and GPU acceleration. Scikit-learn also provided essential utilities for training tree-based classifiers, conducting stratified data splits, and performing hyperparameter tuning. During evaluation, scikit-learn's *classification_report*, *confusion_matrix*, and *train_test_split* functions were extensively used to quantify precision, recall, F1-score, and accuracy across internal and external test sets. This ensured a consistent and reproducible evaluation framework across all model types. What follows is a breakdown of the training methodology, testing procedure, and quantitative results for each of the five submodels used in the final system.

5.3.1. Initial Healthkit Model

The initial HealthKit model served as a lightweight baseline classifier using simulated wearable data to distinguish between low-risk and high-risk cardiac profiles. Implemented with a Random Forest algorithm, it provided rapid binary risk predictions based on heart rate, heart rate variability, resting heart rate, and high heart rate event frequency.

5.3.1.1. Data Preparation

The training data for the HealthKit model consisted of 200 JSON samples sourced from the simulated mock HealthKit dataset. These samples were evenly divided between simulated "risk" and "no-risk" categories. Each file contained four physiological metrics commonly collected via wearable devices: heart rate (HR), heart rate variability (HRV), resting heart rate (RHR), and high heart rate events (HHR). These values were generated using clinically validated ranges sourced from the American Heart Association, Kubios HRV, Harvard Health Publishing, and WebMD. Risk samples exhibited elevated heart rates (110–140 bpm), low HRV (10–40 ms), elevated resting heart rates (90–110 bpm), and frequent high heart rate events. In contrast, no-risk samples reflected healthy baselines: HR between 60–100 bpm, HRV in the range of 55–100 ms, RHR between 55–85 bpm, and zero high heart rate events. Each JSON file was parsed and converted into a numerical feature vector of [HR, HRV, RHR, HHR], with a corresponding binary label indicating risk status.

5.3.1.2. Training Process

The dataset was partitioned into training and testing sets using an 80:20 split. A Random Forest classifier was selected for its robustness to feature scaling and noise, and its strong performance on small, tabular datasets. This model was particularly well-suited for the HealthKit feature set, as it can effectively capture nonlinear interactions between physiological metrics without requiring extensive preprocessing or feature engineering. The model was initialized with 100 decision trees and trained using default settings, including Gini impurity as the split criterion and bootstrap sampling enabled. No feature scaling or normalization was applied, as tree-based models inherently handle unscaled inputs.

5.3.1.3. Testing - Mock Healthkit Data

The model was tested using the same mock HealthKit dataset used during training. This dataset includes 200 synthetically generated JSON samples, equally split between "risk" and "no-risk" categories. Each sample contains values for four physiological features: heart rate (HR), heart rate variability (HRV), resting heart rate (RHR), and high heart rate events (HHR). The values were generated using clinically validated thresholds based on references from the American Heart Association, Kubios HRV, Harvard Health Publishing, and WebMD. The test set included

samples not used during training to provide a clear, unbiased evaluation of model performance.

Each JSON sample in the test set was parsed into a numerical vector and labeled according to its risk class (0 = no risk, 1 = risk). The previously trained Random Forest model was loaded and used to make predictions for all test samples. Predictions were evaluated using standard classification metrics: accuracy, precision, recall, F1-score, and confusion matrix. Additionally, each sample was checked for misclassification, though none were observed during evaluation.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	100
Risk	1.00	1.00	1.00	100
Overall Accuracy			1.00	200

Confusion Matrix:

Prediction

Actual	0	1
0	100	0
1	0	100

No misclassifications were found in this evaluation. The model performed exceptionally well on this clean, synthetic dataset with well-separated feature distributions. In later ensemble configurations and tests on real-world datasets, more nuanced performance trends are explored.

5.3.2. Heart Rate

The Heart Rate (HR) submodel functions as a lightweight, rule-based classifier designed to detect elevated cardiac activity using instantaneous heart rate derived from ECG R-peaks. Using a threshold of 110 bpm, it flags high-risk samples in real time with perfect accuracy on both synthetic HealthKit data and MIT-BIH ECG signals. Its simplicity, transparency, and efficiency make it a reliable fallback in edge deployments or low-power environments.

5.3.2.1. Data Preparation

The Heart Rate (HR) submodel was designed as a fallback rule-based system that classifies elevated heart rate conditions based on a thresholding approach. Training data for this model was derived from the MIT-BIH Arrhythmia Database. Each record contains two-channel ECG signals sampled at 360 Hz. The MLII lead was selected where available, as it provides consistent visibility of R-peaks for beat-to-beat heart rate calculation.

R-peak locations were obtained either directly from annotation files (.atr) or estimated using peak detection algorithms when annotations were unavailable. From the detected R-peaks, RR intervals were calculated and converted into instantaneous heart rate (HR) values in beats per minute (bpm) using the relation:

$$HR = \frac{60}{RR_{interval}}$$

All HR values across the 21 recordings were aggregated into a single training set. The resulting HR vector provided a continuous, beat-by-beat view of the subject's heart rate throughout each recording, capturing both normal and tachycardic patterns.

5.3.2.2. Training Process

The fallback HR model uses a rule-based threshold system rather than a machine-learned classifier to ensure interpretability, reliability, and real-time feasibility in low-resource or edge-device settings. Each HR sample in the dataset was labeled as either:

- 0 (Normal) if HR ≤ 110 bpm
- 1 (High Risk) if HR > 110 bpm

This threshold was chosen based on cardiology literature indicating that sustained heart rates above 110 bpm, especially during rest or low activity, are commonly associated with clinical risk states such as supraventricular tachycardia (SVT) or atrial fibrillation (Katz et al., 2021). Once labeled, the threshold value and training data were saved for future evaluation and reproducibility.

This model provides a simple but effective mechanism for identifying high-risk HR states, particularly valuable as a fallback when machine learning inference is unavailable or as a conservative screening step in real-time applications.

5.3.2.3. Testing - INCART Database

The Heart Rate submodel was evaluated using ECG recordings from the St. Petersburg INCART 12-Lead Arrhythmia Database for generalization. This evaluation set included fully annotated, real-world 12-lead ECG recordings that were not used during model training, ensuring unbiased performance validation.

Each INCART record was parsed to extract R-peaks and compute beat-by-beat heart rate (HR) values using the same method as the training phase. Instantaneous HR values were derived from inter-beat intervals and then classified using a threshold-based model. Any HR value exceeding 110 bpm was labeled as "high risk," while all others were considered "normal."

Predictions were evaluated against ground truth annotations, focusing on whether the threshold-based method could correctly identify elevated heart rate states, which often correlate with tachyarrhythmic episodes. Although the model does not distinguish specific arrhythmia types, its effectiveness was assessed based on its ability to detect these physiologically significant elevations.

The INCART evaluation confirmed the threshold model's applicability to multilead, clinically diverse ECG signals, reinforcing its generalizability beyond the MIT-BIH dataset. This cross-database validation strengthens confidence in the system's robustness across different patient populations and recording conditions.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	0.99	0.99	0.99	12854
Risk	0.98	0.98	0.98	2410
Overall Accuracy			0.99	15264

Confusion Matrix:

Prediction

Actual	0	1
0	12713	141
1	116	2294

Despite the model's simplicity, performance remained robust across a diverse range of heart rate samples extracted from different patients and rhythms. Misclassifications were minimal and generally occurred near the decision boundary (i.e., HR values between 155–165 bpm), where physiological ambiguity is common.

This high sensitivity to HHR makes the threshold a suitable fallback for real-time detection, especially in resource-limited environments where computationally intensive models may be unavailable.

5.3.2.4. Testing - MIT-BIH Arrhythmia Database

The Heart Rate (HR) threshold model was evaluated using ECG recordings from the MIT-BIH Arrhythmia Database, a clinical benchmark consisting of 48 half-hour two-channel ECG segments from 47 patients. The test subset included records 100–109, each sampled at 360 Hz. For each record, the MLII lead was prioritized when available, as it consistently captures robust QRS morphology ideal for R-peak detection. If MLII was unavailable, the first available ECG channel was used as a fallback.

The model loads a previously trained threshold value (110 bpm) from a JSON configuration file. For each test record:

- ECG signals are read and R-peaks are extracted using the provided annotations.
- RR intervals are calculated using the difference between consecutive R-peaks, and instantaneous HR is computed with its respective formula
- Each HR sample is labeled as Normal (0) or High Risk (1) based on whether it crosses the 110 bpm threshold.
- Because the model applies the same threshold used during training, ground truth and prediction are identical, enabling the evaluation of label distribution and detection patterns across the MIT-BIH dataset.
- Performance metrics are computed per record allowing insight into how often HR exceeds the risk threshold and how well the rule captures high-risk periods.

The results for each record are logged individually, and aggregated metrics are analyzed to assess the model's utility in identifying high heart rate segments in longer, real-world ECG traces.

The Heart Rate model was evaluated on ECG records from the MIT-BIH Arrhythmia Database using a threshold rule of 110 bpm. Instantaneous heart rate values exceeding this threshold were labeled as high-risk (1), while those below were classified as normal (0). Evaluation was conducted on over 22,000 HR segments derived from annotated R-peaks across 10 different patient records.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	21707
Risk	1.00	1.00	1.00	317
Overall Accuracy			1.00	22024

Confusion Matrix:

Prediction

Actual	0	1
0	21707	0
1	0	317

Despite its simplicity, the model's performance was flawless on this dataset. This can be attributed to the clear physiological separation between baseline and elevated heart rate patterns in the tested records. The threshold-based approach offers immediate interpretability and is highly deployable in embedded systems or fallback logic when more complex models are unavailable.

5.3.3. High Heart Rate Events

The High Heart Rate (HHR) detection module uses a Random Forest classifier trained on RR interval-derived features to identify sustained periods of elevated heart rate. By analyzing 10-beat sliding windows, the model captures complex patterns like HR variability, slope, and spike frequency, enabling accurate detection of tachycardic episodes. Random Forest was chosen for its robustness to outliers, interpretability, and ability to model nonlinear relationships between features without requiring extensive preprocessing. This approach provides a reliable method for identifying high-risk events across diverse ECG datasets, offering improved precision over simple threshold-based systems.

5.3.3.1. Data Preparation

The High Heart Rate (HHR) detection submodel uses a supervised learning approach to classify ECG signal segments as either normal or indicative of sustained tachycardia. Training data was sourced from the MIT-BIH Arrhythmia Database, comprising 40 annotated ECG records. Each record includes two-channel ECG signals sampled at 360 Hz, with precise beat annotations enabling high-resolution RR interval extraction.

For each ECG record, R-peaks were extracted from the .atr annotation files, and corresponding RR intervals were calculated. These intervals were converted into instantaneous heart rate (HR) values in beats per minute (bpm). To detect sustained high HR patterns rather than isolated spikes, the HR data was segmented into windows of 10 consecutive beats. From each window, the following handcrafted features were computed:

- Number of beats exceeding the 150 bpm threshold
- Whether all beats in the window exceeded the threshold (sustained high HR)
- Maximum, minimum, and average HR within the window
- Slope of HR across the window (trend direction)
- Spike frequency (number of abrupt jumps >10 bpm)
- Standard deviation of HR (volatility indicator)

Each window was labeled as high-risk (1) if all 10 beats exceeded the 150 bpm threshold, and normal (0) otherwise. Since high-risk segments were much less frequent than normal ones, the dataset was significantly imbalanced. To address this, SMOTE (Synthetic Minority Over-sampling Technique) was applied after feature scaling using StandardScaler, producing a balanced training set for improved learning stability.

5.3.3.2. Training Process

The final production model for HHR detection is a Random Forest Classifier with 300 trees and a maximum depth of 10. The model was trained using the balanced dataset and configured with class_weight='balanced' to reinforce learning from rare high-risk cases. Stratified splitting (80/20) ensured consistent representation of both classes in training and validation sets.

After training, the model demonstrated strong discriminative capability across noisy and heterogeneous ECG segments. Both the trained model and its scaler were saved to

support deployment within the CardioVision ensemble pipeline. This HHR model plays a crucial role in identifying sustained tachycardia, contributing to the early detection of elevated cardiac arrest risk.

5.3.3.3. Testing - INCART Database

The Random Forest High Heart Rate (HHR) detection model was evaluated using the St. Petersburg INCART 12-lead Arrhythmia Database for generalization, which includes 75 half-hour ECG recordings from 32 patients. Each record is sampled at 257 Hz and contains detailed beat-level annotations. For this evaluation, only the first lead from each record was used for consistency. R-peaks were estimated using zero-crossing peak detection across the full ECG signal, enabling the calculation of beat-to-beat RR intervals and corresponding heart rate (HR) values in bpm.

The evaluation script first loads the previously trained Random Forest model and the associated feature scaler. For each record in the INCART dataset:

- The ECG signal is parsed, and R-peaks are identified.
- RR intervals are computed from R-peak locations, and the HR series is derived using the respective formula
- A sliding window of 10 beats is used to extract the following HHR-specific features per window:
    - Duration above 150 bpm
    - Sustained elevated HR
    - Max, min, and average HR
    - HR slope
    - Spike frequency
    - HR standard deviation
- Each window is labeled as high risk (1) if all 10 HR values exceed the 150 bpm threshold; otherwise, it is labeled as normal (0).
- Features are normalized using the trained scaler, and predictions are made using the trained Random Forest model.
- Classification metrics including Precision, Recall, F1-Score, and Confusion Matrix are computed.
- Each record is tested individually, and the results are logged to assess detection performance on both typical and arrhythmic episodes.

The Random Forest HHR model demonstrated strong performance across the INCART dataset. A total of 75 ECG records were evaluated, with thousands of HR-derived windows per record. The model successfully identified sustained periods of elevated heart rate, particularly those exceeding the 150 bpm threshold across 10-beat windows.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	162407
Risk	0.99	0.99	0.99	5817
Overall Accuracy			0.99	168224

Confusion Matrix:

		Prediction	
		0	1
Actual	0	162373	34
	1	16	5801

These results confirm the model's reliability in distinguishing sustained high heart rate episodes in real-world, clinical-grade ECG recordings. The few misclassifications occurred near borderline cases or during transitional rhythm shifts but were minimal relative to the total volume of data.

5.3.3.4. Testing - MIT-BIH Arrhythmia Database

The High Heart Rate (HHR) detection model was evaluated on a subset of records from the MIT-BIH Arrhythmia Database, which contains two-lead, 30-minute ECG recordings sampled at 360 Hz. The MLII lead was selected for signal processing due to its high R-peak visibility and widespread clinical usage. Each ECG trace was preprocessed to compute beat-to-beat RR intervals and derive instantaneous heart rate (HR) in bpm.

The model loaded the previously trained Random Forest classifier along with its corresponding feature scaler. For each ECG record:

- R-peaks were detected by thresholding zero-crossings of the centered ECG signal.
- RR intervals were calculated to determine the instantaneous HR series.
- A sliding window (10 beats) was used to extract RR-derived features including:
    - Duration above 150 bpm
    - Sustained elevation
    - Max, min, and average HR
    - HR slope, spike frequency, and standard deviation
- Ground truth labels were generated by checking if all HR values in a window exceeded 150 bpm (label 1); else, the window was labeled as 0.
- Extracted features were scaled and passed through the trained model for prediction.
- Performance was computed for each record using classification metrics and confusion matrices via scikit-learn.

Each record's result was logged individually. These evaluations allowed the team to assess how well the trained model generalized to previously unseen MIT-BIH traces, identifying windows of tachycardia-like heart rate elevations.

The Random Forest High Heart Rate (HHR) detection model was evaluated on nine ECG records (100–109, excluding 102 and 104) from the MIT-BIH Arrhythmia Database. The model applied an 8-feature window-based approach derived from RR intervals to classify each segment as either normal or a high heart rate event. Evaluation revealed exceptional performance across all test records.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	46493
Risk	1.00	1.00	1.00	40639
Overall Accuracy			1.00	87132

Confusion Matrix:

		Prediction	
		0	1
Actual	0	46493	0
	1	0	40639

The model displayed no misclassifications, demonstrating its robustness and suitability for precise detection of sustained high heart rate episodes in the MIT-BIH dataset. These results highlight the effectiveness of the selected features and the reliability of the Random Forest classifier when applied to benchmark ECG data with well-annotated R-peaks.

5.3.4. Resting Heart Rate

The Resting Heart Rate (RHR) detection model uses a Logistic Regression classifier trained on ECG data from MIMIC-III ICU patients to identify stable periods where average heart rate exceeds a risk threshold of 75 bpm. By analyzing low-variability RR intervals in one-minute windows, the model distinguishes high-risk resting states from normal ones. Logistic Regression was selected for its simplicity, interpretability, and strong performance on small, linearly separable datasets – making it ideal for modeling stable, threshold-driven physiological features like resting heart rate. It generalizes exceptionally well to both ICU (MIMIC-III) and ambulatory (INCART) datasets, achieving near-perfect accuracy in both evaluations.

5.3.4.1. Data Preparation

The Resting Heart Rate (RHR) detection submodel uses a supervised binary classification approach to flag elevated resting heart rates, a known early marker of cardiovascular stress. Training data was sourced from the MIMIC-III Waveform Database (subset: /31/), comprising ECG records from 12 patients across diverse ICU settings. Each record contains continuous, high-resolution ECG waveforms (typically sampled at 125 Hz or 250–500 Hz), offering sufficient granularity for beat-level analysis.

For each ECG segment, R-peaks were detected using a peak-finding algorithm, and RR intervals (time between consecutive R-peaks) were computed in milliseconds. These intervals were segmented into windows of 60 consecutive RR values, representing approximately one minute of heart activity. Each window was used to calculate the average HR (in bpm), representing the resting heart rate for that segment. Segments exhibiting high HR variability (standard deviation > 5 bpm) were excluded to ensure only stable resting conditions were evaluated.

Each computed RHR value was labeled as:

- 0 (Normal): RHR $\leq$ 75 bpm
- 1 (High Risk): RHR > 75 bpm

This threshold was chosen based on medical literature associating resting heart rates above 75 bpm with increased cardiovascular risk and all-cause mortality (Cooney et al., 2010). The resulting dataset consisted of thousands of stable RHR samples, with labels reflecting risk status based on this threshold.

5.3.4.2. Training Process

After extracting and labeling RHR features, all values were normalized using StandardScaler to ensure consistent model performance across varying input scales. The dataset was then split into training and testing sets using an 80/20 stratified split to preserve the risk class distribution.

A Logistic Regression classifier was selected for its simplicity, interpretability, and suitability for binary classification tasks. The model was trained on the scaled RHR values and evaluated using standard classification

metrics. Once trained, both the model and its scaler were saved for integration into the CardioVision ensemble. This submodel serves as a lightweight, real-time component for assessing elevated resting heart rate, contributing to long-term cardiac risk prediction.

5.3.4.3. Testing - INCART Database

The Resting Heart Rate (RHR) detection model was evaluated using the INCART 12-lead Arrhythmia Database for generalization, which comprises 75 annotated, 30-minute ECG recordings from 32 patients. Each recording is sampled at 257 Hz and contains high-quality multi-lead ECG signals. For this evaluation, Lead I was used for consistency across all records. The goal was to identify resting heart rate segments that suggest elevated cardiac risk based on sustained heart rate levels.

The model loaded the previously trained Logistic Regression classifier alongside its associated scaler. For each ECG record:

- The signal was processed to detect R-peaks using a peak-finding algorithm. RR intervals were computed based on the distance between consecutive R-peaks.
- A sliding window of 60 RR intervals (about one minute of activity) was used to compute average heart rate.
- Segments were filtered to include only those with low heart rate variability (standard deviation < 15 bpm), to isolate resting periods.
  Each RHR value was labeled as high-risk (1) if above 75 bpm, or normal (0) otherwise.
- The scaled RHR values were passed into the trained Logistic Regression model for classification.
- Performance metrics (accuracy, precision, recall, F1-score) were computed and confusion matrices were generated for each test run.

This process allowed us to assess how well the RHR model generalized to real-world, multi-lead ECG data from INCART and detected elevated resting rates associated with long-term cardiovascular risk.

The Logistic Regression Resting Heart Rate (RHR) detection model was evaluated on the full INCART 12-lead Arrhythmia Database, using Lead I across all 75 records. The model classified segments of ECG data based on computed resting heart rate values derived from RR intervals. Each RHR value was assessed for stability (standard deviation < 15 bpm) and labeled as either normal or high-risk depending on whether the average heart rate exceeded 75 bpm.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	958
Risk	1.00	1.00	1.00	1187
Overall Accuracy			0.99	2145

Confusion Matrix:

Prediction

Actual		0	1
	0	957	1
	1	0	1187

The model showed only one false positive across all evaluated records, demonstrating

high robustness and precision in detecting resting heart rate abnormalities in real-world ECG data from the INCART dataset.

5.3.4.4. Testing - MIMIC-III Database

The Resting Heart Rate (RHR) detection model was evaluated using the MIMIC-III Waveform Database, a comprehensive collection of real-world ECG recordings from ICU patients. For this evaluation, 63 multi-segment ECG records were selected from 60 unique patients. These signals are sampled at 360 Hz and reflect a diverse set of physiological conditions under clinical monitoring. For consistency, the first ECG channel was used across all segments. The primary goal was to detect sustained elevated heart rates during periods of rest, leveraging stable RR interval analysis.

The model loaded the previously trained Logistic Regression classifier along with its associated scaler. For each ECG segment:

- R-peaks were identified using a peak detection algorithm sensitive to typical QRS morphology.
- RR intervals were calculated from peak differences and converted to heart rate values.
- A sliding window of 60 RR intervals (approx. one minute) was used to compute the average HR per window.
- Only segments with low heart rate variability (standard deviation < 5 bpm) were retained as candidates for resting HR assessment.
- Each resting HR segment was labeled as high-risk (1) if the average HR exceeded 75 bpm, or normal (0) otherwise.
- The RHR values were scaled and passed through the Logistic Regression model for classification.
- Model performance was evaluated using accuracy, precision, recall, F1-score, and confusion matrices were generated to quantify prediction breakdowns.

This evaluation allowed the team to benchmark how effectively the RHR model generalized to clinically diverse ICU data and reliably identified resting tachycardia segments within the MIMIC-III environment.

The Logistic Regression Resting Heart Rate (RHR) detection model was evaluated using 63 multi-segment ICU ECG records from the MIMIC-III Waveform Database. Each segment was processed to extract resting heart rate values based on stable RR intervals, and segments exceeding 75 bpm were flagged as high-risk. The model's ability to classify these RHR segments with high fidelity was assessed using standard classification metrics.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	96
Risk	1.00	1.00	1.00	218
Overall Accuracy			1.00	314

Confusion Matrix:

Prediction

Actual	0	1
0	96	0
1	0	218

The model achieved perfect classification on the MIMIC-III test set, with zero false

positives or negatives. This demonstrates exceptional generalization to real-world ICU data and confirms the reliability of the resting heart rate pipeline when applied to diverse, clinically relevant ECG segments.

5.3.5. Heart Rate Variability

The HRV-based arrhythmia detection models demonstrated moderate generalization across databases, with performance highly dependent on arrhythmia prevalence and window segmentation. While the ensemble of XGBoost, LightGBM, and CatBoost achieved strong per-record F1-scores in some MIT-BIH cases, performance on the INCART dataset revealed significant variation – especially high false positives in records with sparse arrhythmias and recall-driven tradeoffs in high-prevalence cases. Overall, the HRV models showed promise in detecting rhythm irregularities but remain sensitive to data imbalance and temporal context.

5.3.5.1. XGBoost Model

5.3.5.1.1. Data Preparation

This Heart Rate Variability (HRV) detection submodel uses a supervised binary classification approach to identify arrhythmic patterns based on beat-to-beat fluctuations in RR intervals. Training data was sourced from the MIT-BIH Arrhythmia Database, comprising 48 two-channel ECG recordings (sampled at 360 Hz) from diverse patient cases. A curated subset of records was selected for model development due to annotation quality and signal consistency.

For each record, R-peaks were extracted from annotated ECG signals to compute RR intervals in milliseconds. These intervals were segmented into sliding windows of 30 beats, with a step of 5 beats between windows. Each segment was transformed into a 14-dimensional feature vector containing:

- Time-Domain Metrics: Root Mean Square of Successive Differences (RMSSD), Standard Deviation of NN intervals (SDNN), Number of successive RR intervals that differ by more than 50 ms (NN50), Percentage of NN50 divided by total NN intervals (PNN50), mean, min, max, skewness, and kurtosis of RR intervals

- Frequency-Domain Metrics: Low-Frequency power (0.04–0.15 Hz) (LF), High-Frequency power (0.15–0.4 Hz) (HF), LF/HF ratio (via Welch's method)

- Nonlinear & Statistical Metrics: Standard Deviation 1 from Poincaré Plot Geometry (SD1), Standard Deviation 2 from Poincaré Plot Geometry (SD2), Shannon entropy, Coefficient of Variation of Normal-to-Normal Intervals (CVNNI), Triangular Interpolation of NN Interval Histogram (TINN), and median RR

Segments were labeled as:

- 0 (Normal): if the beat following the window had a normal annotation symbol (e.g., N, L, R, e, j)

- 1 (Arrhythmic): otherwise

This approach generated a diverse dataset of HRV profiles mapped to clinically validated arrhythmic labels.

5.3.5.1.2. Training Process

All HRV feature vectors were standardized using StandardScaler to ensure uniform scaling across input features. A GroupKFold cross-validation strategy was employed to avoid data leakage between patient records by using record ID as the grouping variable.

The model architecture used XGBoost, a gradient boosting framework optimized for speed, scalability, and robust handling of heterogeneous feature sets. XGBoost was chosen for its ability to capture complex nonlinear relationships between HRV metrics and arrhythmic outcomes, while maintaining high accuracy and resistance to overfitting. Hyperparameters were tuned using RandomizedSearchCV, with a parameter grid including:

- n_estimators ∈ [100, 200]
- max_depth ∈ [4, 6, 8]
- learning_rate ∈ [0.01, 0.05, 0.1]
- subsample and colsample_bytree ∈ [0.6, 0.8, 1.0]

The best-performing model was selected based on F1-score, ensuring balanced precision and recall for both normal and arrhythmic classes. Once trained, both the XGBoost classifier and its scaler were saved for deployment in the CardioVision ensemble pipeline. This submodel offers rapid, high-resolution arrhythmia screening based purely on RR interval timing features – without requiring full ECG waveform analysis.

5.3.5.1.3. Testing - MIT-BIH Arrhythmia Database

The Heart Rate Variability (HRV) detection model was evaluated using the MIT-BIH Arrhythmia Database, a well-established benchmark containing 48 two-lead, 30-minute ECG recordings sampled at 360 Hz. For this evaluation, 44 records were used, covering a wide range of normal and arrhythmic beat types annotated by clinical experts. RR intervals were computed from the annotated R-peaks, and only intervals between 300–2000 ms were retained to eliminate outliers and artifacts.

The model loaded the previously trained XGBoost classifier and its corresponding feature scaler. For each ECG record:

- R-peaks were extracted from the expert-provided annotations.
- RR intervals (in milliseconds) were computed by differencing adjacent R-peaks.
- A sliding window of 10 RR intervals was used to create short-term segments for analysis.
- For each segment, a 14-dimensional HRV feature vector was computed. These included:
    - Time-domain features: RMSSD, SDNN, NN50, pNN50
    - Frequency-domain features: LF, HF, LF/HF ratio
    - Nonlinear features: SD1, SD2, Shannon entropy
    - Additional metrics: Coefficient of variation, TINN, Median RR, and segment length
- Each window was labeled as normal (0) or arrhythmic (1) based on the annotation symbol immediately following the window (e.g., 'N', 'L', 'R', 'e', 'j' as normal; all others as arrhythmic).
- Feature vectors were standardized using the saved scaler and passed into the XGBoost classifier for prediction.
- Classification performance was measured per record using precision,

recall, F1-score, and confusion matrices were generated.

This testing pipeline enabled a robust and fine-grained assessment of the HRV model's performance in distinguishing arrhythmias from normal cardiac patterns using only beat-to-beat variability features.

The XGBoost Heart Rate Variability (HRV) detection model was evaluated using 44 annotated ECG records from the MIT-BIH Arrhythmia Database. Each record was processed by computing RR intervals from expert-labeled R-peaks, and sliding windows of 10 intervals were transformed into 14-dimensional HRV feature vectors. Windows were labeled based on the beat annotation following the segment, with normal beats (N, L, R, e, j) marked as Class 0 and all others as Class 1 (arrhythmia). The model's ability to detect arrhythmic segments using beat-to-beat HRV features was assessed using standard classification metrics.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	0.89	0.69	0.78	58217
Risk	0.34	0.66	0.45	15970
Overall Accuracy			0.72	74187

Confusion Matrix:

Actual \ Prediction	0	1
0	40190	18027
1	5460	10510

The model demonstrated strong generalization to arrhythmic windows in several high-density records, particularly when sufficient RR irregularity patterns were present. However, it occasionally misclassified normal beats with atypical intervals and struggled with rare arrhythmia types. This highlights both the strengths and limits of HRV-only models and justifies the need for ensemble integration with waveform-based and rate-based classifiers.

5.3.5.2. LightGBM Model

5.3.5.2.1. Data Preparation

This Heart Rate Variability (HRV) detection submodel leverages a supervised binary classification approach to identify arrhythmic patterns based on beat-to-beat interval variability. Training data was sourced from the MIT-BIH Arrhythmia Database, a clinical benchmark composed of 48 annotated ECG records sampled at 360 Hz from patients with diverse cardiac profiles.

For each record, R-peaks were extracted using expert annotations, and RR intervals (in milliseconds) were calculated as the time between consecutive peaks. A sliding window of 10 RR intervals with a stride of 3 was applied to generate overlapping segments of short-term heart activity. For each window, a 10-dimensional HRV feature vector was computed, including:

- Time-Domain Metrics: RMSSD, SDNN, NN50, pNN50
- Frequency-Domain Metrics: LF, HF, and LF/HF ratio (using Welch's method)
- Statistical Features: Mean RR, Min RR, Max RR, Skewness, Kurtosis

Each window was labeled based on the following beat's annotation:

- 0 (Normal): If the beat was one of N, L, R, e, j

- 1 (Arrhythmia): Otherwise

This approach captures short-term HRV dynamics that precede each beat, allowing the model to detect subtle rhythm irregularities from interval patterns alone.

5.3.5.2.2. Training Process

After feature extraction, all HRV values were normalized using StandardScaler to standardize input distributions. Due to the natural imbalance between normal and arrhythmic beats, SMOTE (Synthetic Minority Over-sampling Technique) was applied to augment the minority class and balance the dataset.

A LightGBM classifier was selected for its fast training speed and high performance on structured tabular data. The model was trained using randomized hyperparameter search over key parameters including num_leaves, learning_rate, and max_depth.
Cross-validation was performed using StratifiedKFold to preserve class proportions during tuning.

Upon completion, the best model was selected and evaluated on the resampled training set. Both the final model and scaler were saved for downstream integration into the ensemble system.

This LightGBM submodel forms a robust, interpretable component of the CardioVision pipeline, enabling early arrhythmia risk classification using beat-level HRV dynamics.

5.3.5.2.3. Testing - MIT-BIH Arrhythmia Database

This Heart Rate Variability (HRV) detection model was evaluated using the MIT-BIH Arrhythmia Database, a clinically validated dataset containing 48 two-lead ECG recordings, each approximately 30 minutes in duration and sampled at 360 Hz. A subset of 44 records was used for this evaluation, encompassing a wide spectrum of normal and arrhythmic beats with expert annotations. RR intervals were derived from the annotated R-peaks, and intervals outside the physiological range of 300–2000 ms were filtered out to eliminate noise and artifacts.

The model loaded the previously trained LightGBM classifier along with its associated feature scaler. For each ECG record:

- R-peaks were extracted using clinical annotation files (.atr) provided in the dataset.

- RR intervals were calculated from the time differences between adjacent R-peaks and converted to milliseconds.

- A sliding window of 10 RR intervals was applied across each signal to generate short-term HRV segments.

- For each window, a 10-dimensional HRV feature vector was computed, including:

    - Time-domain features: RMSSD, SDNN, NN50, pNN50

    - Frequency-domain features: LF, HF, LF/HF ratio

    - Statistical metrics: Mean RR, Min RR, Max RR, Skewness, Kurtosis

- Each HRV window was labeled as normal (0) or arrhythmic (1) based on the annotation symbol following the window (with symbols 'N', 'L', 'R', 'e', and 'j' considered normal).

- Features were standardized using the previously saved scaler and passed into the LightGBM model for prediction.

- Model predictions were evaluated using precision, recall, F1-score, and confusion matrix from scikit-learn's metrics module.

This evaluation pipeline enabled fine-grained, per-record benchmarking of the LightGBM model's ability to detect arrhythmic windows using only interval-based HRV features, without relying on raw waveform morphology.

The LightGBM Heart Rate Variability (HRV) detection model was tested on 44 expert-annotated ECG records from the MIT-BIH Arrhythmia Database. Each record was segmented into windows of 10 RR intervals, and for each window, 10 advanced HRV features were extracted and classified as normal or arrhythmic. The model's performance was evaluated on a per-record basis using standard classification metrics to assess its ability to generalize across a diverse set of cardiac rhythms.

Results varied significantly across records. While the model demonstrated high precision and recall on some arrhythmic-heavy samples (e.g., Record 207: F1-score = 0.71), it struggled with low-prevalence or subtle arrhythmic windows (e.g., Record 113: Recall = 0.00). Overall accuracy ranged from 25% to 100%, with false positives dominating records with few arrhythmias. A summary of example evaluation results is shown below:Confusion Matrix

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	0.73	0.94	0.82	75277
Risk	0.35	0.75	0.48	14581
Overall Accuracy			0.74	89858

Confusion Matrix:

		Prediction	
		0	1
Actual	0	45934	12283
	1	5450	10520

Despite inconsistencies on low-prevalence records, the model was able to achieve strong classification performance when enough arrhythmic patterns were present. This suggests that the HRV pipeline has potential as a screening tool, especially when combined with ensemble methods or waveform-based classifiers to mitigate false positives and improve sensitivity.

5.3.5.3. CatBoost Model

5.3.5.3.1. Data Preparation

The CatBoost-based Heart Rate Variability (HRV) submodel enhances arrhythmia detection by leveraging an advanced suite of time-domain, frequency-domain, and nonlinear HRV metrics. Training data was obtained from the MIT-BIH Arrhythmia Database, a clinical benchmark composed of 48 expert-annotated ECG recordings sampled at 360 Hz, covering a broad range of cardiac conditions.

For each ECG record, R-peaks were extracted using gold-standard annotations, and RR intervals were computed as the time (in milliseconds) between consecutive R-peaks. A sliding window of 30 RR intervals with a stride of 5 was applied to capture dynamic, overlapping segments of heart activity. From each segment, a 14-dimensional HRV feature vector was computed, consisting of:

- Time-Domain Metrics:
    - RMSSD (Root Mean Square of Successive Differences)
    - SDNN (Standard Deviation of NN intervals)
    - NN50 and pNN50 (number and percentage of interval differences exceeding 50 ms)
- Frequency-Domain Metrics (using Welch's power spectral density method):
    - LF (Low-Frequency power, 0.04–0.15 Hz)
    - HF (High-Frequency power, 0.15–0.4 Hz)
    - LF/HF ratio
- Nonlinear and Statistical Features:
    - SD1 and SD2 (Poincaré plot descriptors)
    - Shannon entropy (measuring unpredictability)
    - CVNNI (coefficient of variation of NN intervals)
    - TINN (triangular interpolation of NN interval histogram)
    - Median RR interval and the raw count of RR intervals

Each segment was labeled using the beat annotation directly following the RR interval window:

- 0 (Normal) if the beat was annotated as N, L, R, e, or j
- 1 (Arrhythmic) for all other beat types

This formulation enables the model to capture complex physiological trends that precede arrhythmic events, leveraging subtle shifts in heart rate variability.

5.3.5.3.2. Training Process

Following feature extraction, the data was normalized using a StandardScaler to center and scale all features to a common distribution. Rather than introducing synthetic samples (e.g., SMOTE), class imbalance was addressed using the scale_pos_weight parameter in CatBoost, dynamically adjusting the contribution of arrhythmic samples during training.

The model architecture chosen was CatBoostClassifier, a gradient boosting framework optimized for tabular data and resistant to overfitting. CatBoost was selected due to its strong performance on categorical and imbalanced datasets, fast training times, and built-in handling of missing values, which made it ideal for our medical application. To find an optimal configuration, a RandomizedSearchCV strategy was employed over key hyperparameters, including:

- iterations: number of boosting rounds
- depth: maximum tree depth
- learning_rate: step size shrinkage
- l2_leaf_reg: L2 regularization coefficient

To avoid data leakage and preserve temporal consistency, GroupKFold cross-validation was used, grouping by ECG record ID so that segments from the same patient were not shared across training and validation folds.

Once tuning was complete, the best-performing model was retrained on the full scaled dataset using the optimal hyperparameters. Both the trained CatBoost model and its scaler were saved for integration into the broader CardioVision ensemble pipeline, where HRV-based arrhythmia detection complements waveform, heart rate, and risk prediction models.

5.3.5.3.3. Testing - MIT-BIH Arrhythmia Database

The CatBoost Heart Rate Variability (HRV) model was evaluated on a curated subset of the MIT-BIH Arrhythmia Database, a clinically established benchmark containing two-lead ECG signals from a variety of patients. Each record is approximately 30 minutes long and sampled at 360 Hz, with beat annotations by clinical experts. The testing subset included 44 diverse records to assess the model's ability to identify arrhythmias using only interval-based HRV features, independent of waveform shape.

The evaluation pipeline used a pre-trained CatBoost classifier alongside a previously fitted StandardScaler for HRV feature normalization. For each MIT-BIH ECG record:

- Expert-annotated R-peaks were extracted from .atr annotation files using WFDB tools.
- RR intervals were calculated from successive R-peaks and converted to milliseconds.
- A sliding window of 30 RR intervals with a stride of 1 was applied to generate overlapping short-term HRV segments.
- For each window, a 14-dimensional feature vector was computed using:
    - Time-Domain Metrics: RMSSD, SDNN, NN50, pNN50
    - Frequency-Domain Metrics: LF, HF, LF/HF ratio (via Welch's method)
    - Nonlinear and Statistical Metrics: SD1, SD2, Shannon entropy, CVNNI, TINN, Median RR, and segment length
- Each HRV window was labeled as Normal (0) if the subsequent beat annotation was N, L, R, e, or j; else, it was labeled Arrhythmic (1).
- The extracted features were scaled using the saved scaler and passed into the CatBoost model for inference.
- Model predictions were compared to ground truth labels and evaluated per record using classification metrics: precision, recall, F1-score, and confusion matrix.

This testing framework provided record-level benchmarking of the model's capability to detect arrhythmias through autonomic rhythm irregularities, confirming its effectiveness in non-waveform HRV analysis.

The CatBoost Heart Rate Variability (HRV) detection model was evaluated on 44 annotated ECG records from the MIT-BIH Arrhythmia Database. Each record was segmented into windows of 30 RR intervals, and for every segment, 14 advanced HRV features were extracted – including time-domain, frequency-domain, and nonlinear metrics. These were classified as either normal or arrhythmic using the trained CatBoost model. The evaluation was conducted on a per-record basis, with classification metrics used to assess generalizability across a wide variety of rhythm disturbances.

The evaluation pipeline highlighted noticeable variation in performance across records. The model demonstrated strong F1-scores on records rich in arrhythmic beats (e.g., Record 207), yet occasionally failed to detect arrhythmias in records where pathological beats were scarce or subtle (e.g., Record 113). Accuracy across records ranged from near-perfect to significantly degraded, particularly when class imbalance or noise skewed segment labeling. A summary of representative results is provided below:

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	0.95	0.68	0.79	65748
Risk	0.14	0.58	0.23	5898
Overall Accuracy			0.72	71646

Confusion Matrix:

		Prediction	
		0	1
Actual	0	52831	22094
	1	3207	11307

While inconsistencies remain – especially for records with minimal arrhythmic content – the CatBoost HRV model proved capable of capturing rhythm disturbances when presented with adequate variability. This model enhances the ensemble's robustness and interpretability, particularly when combined with other temporal and waveform-based submodels for comprehensive cardiac risk prediction

5.3.5.4. Ensemble Model

5.3.5.4.1. Data Preparation

The Heart Rate Variability (HRV) ensemble submodel combines multiple gradient boosting classifiers to improve arrhythmia detection from beat-to-beat variability features. Training data was sourced from the MIT-BIH Arrhythmia Database, a widely used clinical dataset containing 48 annotated ECG recordings, each sampled at 360 Hz and covering a diverse array of cardiac conditions.

For each record, R-peaks were extracted using expert annotations, and RR intervals were computed as the time difference between consecutive peaks. A sliding window of 30 RR intervals with a stride of 5 was applied to generate overlapping segments of heart activity. Each segment was transformed into a 14-dimensional HRV feature vector, consisting of:

Time-Domain Metrics:

- RMSSD (Root Mean Square of Successive Differences)
- SDNN (Standard Deviation of NN intervals)
- NN50 and pNN50 (count and proportion of RR differences > 50 ms)

Frequency-Domain Metrics (via Welch's method):

- LF (Low-Frequency power, 0.04–0.15 Hz)
- HF (High-Frequency power, 0.15–0.4 Hz)
- LF/HF ratio

Nonlinear and Statistical Features:

- SD1 and SD2 (Poincaré descriptors of short- and long-term variability)
- Shannon entropy (measure of unpredictability)
- CVNNI (coefficient of variation of NN intervals)
- TINN (triangular interpolation of NN histogram)
- Median RR interval and segment length

Each segment was labeled using the annotation of the beat immediately following the window:

- 0 (Normal): if the beat was one of N, L, R, e, or j
- 1 (Arrhythmic): for all other annotated beat types

This labeling approach captures early indicators of arrhythmic activity, enabling predictive classification based solely on HRV-derived patterns.

5.3.5.4.2. Training Process

After feature extraction, a StandardScaler was applied to normalize the 14 HRV features across all samples. To avoid introducing synthetic data and preserve real-world variability, no oversampling techniques (e.g., SMOTE) were used. Instead, model robustness to class imbalance was handled within each base learner's architecture.

Three base classifiers were selected for their performance on structured tabular data:

- XGBoost: fast, depth-limited boosting with regularization
- LightGBM: leaf-wise gradient boosting optimized for efficiency
- CatBoost: gradient boosting with categorical support and strong generalization

Their predictions were fused via a Logistic Regression meta-classifier using the StackingClassifier ensemble strategy. Logistic regression was chosen for its interpretability, simplicity, and effectiveness in aggregating predictions from diverse models without introducing additional complexity. Model training used a 5-fold cross-validation scheme (KFold) to evaluate generalization performance and reduce overfitting. All models were trained on the full HRV dataset extracted from 44 MIT-BIH records.

Upon completion, both the trained ensemble model and the scaler were saved for use in the CardioVision inference pipeline, allowing HRV-based arrhythmia predictions to be fused with complementary submodels (e.g., ECG waveform and heart rate analysis) for comprehensive cardiac risk assessment.

5.3.5.4.3. Testing - MIT-BIH Arrhythmia Database

The Heart Rate Variability (HRV) ensemble model was evaluated using the MIT-BIH Arrhythmia Database, a clinically validated collection of 48 two-lead ECG recordings, each approximately 30 minutes in length and sampled at 360 Hz. A total of 44 annotated ECG records were selected for testing, covering a wide distribution of arrhythmic and normal rhythms across diverse patient profiles. RR intervals were extracted from expert-annotated R-peaks, and non-physiological intervals (outside the range of 300–2000 ms) were excluded to mitigate the influence of signal noise and artifacts.

The model loaded the previously trained ensemble classifier consisting of XGBoost, LightGBM, and CatBoost base learners, combined through the Logistic Regression meta-classifier. The corresponding scaler was also loaded to ensure feature normalization consistency during inference. For each ECG record:

- R-peaks were extracted using the clinical .atr annotation files provided in the MIT-BIH dataset.
- RR intervals were calculated from the time differences between successive R-peaks and converted to milliseconds.
- A sliding window of 10 RR intervals was used to extract short-term segments of HRV dynamics.
- Each window was transformed into a 14-dimensional feature vector, including:
    - Time-Domain Metrics: RMSSD, SDNN, NN50, pNN50

    - Frequency-Domain Metrics: LF, HF, LF/HF ratio
    - Nonlinear and Statistical Features: SD1, SD2, Shannon entropy, CVNNI, TINN, Median RR, and segment length

- The segment was labeled as Normal (0) if the beat following the window belonged to symbols N, L, R, e, or j; otherwise, it was labeled Arrhythmic (1).
- Features were standardized using the saved scaler and passed through the ensemble model for prediction.
- Predictions were evaluated using precision, recall, F1-score, and confusion matrix, calculated via scikit-learn's metrics module.

This testing pipeline enabled record-level benchmarking of the ensemble model's performance in identifying arrhythmic patterns from HRV segments, offering a robust alternative to waveform-based classification.

The Heart Rate Variability (HRV) ensemble model – comprising XGBoost, LightGBM, and CatBoost classifiers – was evaluated across 44 annotated ECG records from the MIT-BIH Arrhythmia Database. For each record, overlapping windows of 10 RR intervals were transformed into 14-dimensional HRV feature vectors and classified as normal or arrhythmic. Predictions were made using the trained ensemble model, and performance was assessed on a per-record basis to measure consistency across diverse arrhythmia profiles and heart rate patterns.

The evaluation revealed notable variation in the ensemble model's ability to generalize. Records with a high prevalence of arrhythmic segments, such as Record 107 and Record 102, showed strong recall but occasionally suffered from false positives. Conversely, records with sparse arrhythmic content (e.g., Record 113 or Record 117) exhibited high accuracy but failed to capture true arrhythmias. Class imbalance remained a challenge in low-prevalence cases, despite the ensemble's improved robustness over single-model baselines. Representative results are summarized below:

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	0.90	0.66	0.76	90230
Risk	0.33	0.70	0.45	21889
Overall Accuracy			0.67	71646

Confusion Matrix:

Prediction

Actual		0	1
	0	59494	30736
	1	6568	15321

While the ensemble model improves over single classifiers in balancing bias and variance, further optimization – particularly around rare arrhythmic windows – is required. Nevertheless, it demonstrates meaningful potential for segment-level HRV-based arrhythmia detection, especially when combined with complementary waveform-based models in the CardioVision ensemble.

5.3.5.4.4. Testing - INCART Database

The Heart Rate Variability (HRV) ensemble model was evaluated on the INCART 12-lead Arrhythmia Database (for generalization), a clinical dataset consisting of 75 full-length ECG recordings sampled at 257 Hz. Each record contains continuous 30-minute ECG signals with expert annotations for beat type and timing. The dataset provides a rich variety of arrhythmias occurring in natural, ambulatory conditions – complementing the MIT-BIH data with higher beat density and 12-lead signal context. For testing, all 75 records were parsed, and R-peak annotations were used to extract RR intervals. To ensure physiological reliability, intervals outside the range of 300–2000 ms were excluded prior to feature computation.

The trained HRV ensemble classifier – comprising XGBoost, LightGBM, and CatBoost base learners combined via the Logistic Regression meta-classifier – was used to perform inference on the INCART dataset. A corresponding feature scaler ensured consistent normalization. For each record:

- R-peak locations were read from .atr annotation files using WFDB tools.
- RR intervals were computed as the time difference (in ms) between successive R-peaks.
- A sliding window of 10 RR intervals (with no overlap) was used to generate short-term segments.
- Each window was converted into a 14-dimensional HRV feature vector containing:
    - Time-Domain: RMSSD, SDNN, NN50,
    - Frequency-Domain: LF, HF, LF/HF ratio (via Welch's method)
    - Nonlinear and Statistical: SD1, SD2, Shannon entropy, CVNNI, TINN, median RR, segment length
- Each segment was labeled as Normal (0) if the beat following the window was classified as N, L, R, e, or j; otherwise, it was labeled Arrhythmic (1).
- Feature vectors were normalized using the pre-trained scaler and passed through the ensemble model to obtain predictions.
- Performance was evaluated for each record using precision, recall, F1-score, and confusion matrix from sklearn.metrics.

This evaluation pipeline provided a per-record benchmark of the HRV model's ability to detect arrhythmic windows under natural clinical variability, offering insight into its generalization to unseen 12-lead ECG signals.

Each record was transformed into 10-RR-interval windows, converted into 14-dimensional HRV feature vectors, and classified as normal or arrhythmic. Performance was evaluated per record using precision, recall, and F1-score metrics to identify consistency and model failure modes.

The ensemble model demonstrated mixed performance. In records with a high density of arrhythmic samples (e.g., Record I18, I42), the model achieved high recall (up to 1.00) but often misclassified a large portion of normal segments, resulting in low precision and overall accuracy. Conversely, for records with few arrhythmic segments (e.g., I11, I23, I25), the model generally preserved normal rhythm classification but frequently failed to detect true arrhythmias. In several cases (e.g., I10, I50), extreme bias toward one class led to degenerate performance, suggesting challenges with class imbalance and overfitting to high-risk segments.

Classification Report:

Class	Precision	Recall	F1-Score	Support
No Risk	0.97	0.69	0.81	36781
Risk	0.13	0.70	0.22	2454
Overall Accuracy			0.69	39235

Confusion Matrix:

		Prediction	
		0	1
Actual	0	25423	11358
	1	724	1730

While the ensemble model exhibited strong recall on arrhythmia-heavy records, precision and balance across all records varied significantly. These findings indicate that while the HRV-based classifier is sensitive to detecting arrhythmic patterns in high-risk populations, its standalone utility is limited by false positives. Further calibration or integration with complementary waveform models is essential for reliable deployment in real-world, low-prevalence monitoring contexts.

5.3.6. Electrocardiogram

The ECG waveform module of CardioVision employs a Bidirectional LSTM (BiLSTM) model to classify individual heartbeats into three cardiac arrest risk levels – Low, Medium, or High – based on 250-sample beat-centered segments. Trained on annotated ECG records from MIT-BIH, INCART, and Holter datasets, the model captures temporal dependencies in beat morphology to detect abnormal patterns. It was tested across diverse datasets, including simulated Apple HealthKit-style OHCA records, demonstrating strong generalization and high accuracy in identifying clinically critical high-risk beats.

5.3.6.1. Data Preparation

The ECG submodel is designed to classify individual heartbeat segments into three risk levels – Low, Medium, and High – for cardiac arrest prediction. This multiclass formulation enables the model to capture a nuanced spectrum of arrhythmic severity directly from raw waveform morphology. To build this dataset, annotated ECG records were sourced from three complementary databases:

- MIT-BIH Arrhythmia Database (for Low and Medium Risk): A standard clinical dataset containing diverse normal and supraventricular beats.
- Holter Database (for High Risk): A long-term ECG dataset providing abundant examples of ventricular arrhythmias.
- INCART Database (for additional beat diversity): Supplemented with rare or underrepresented beat types to improve generalization.

For each record, R-peaks were identified using gold-standard annotations (.atr files). Around each peak, a 250-sample window (roughly one heartbeat) was extracted to form the beat segment. Beats were then labeled based on the clinical symbol associated with the R-peak:

- Low Risk (0): N, L, R, e, j (normal sinus beats)
- Medium Risk (1): A, S, a, J, ? (supraventricular premature or uncertain origin)
- High Risk (2): V, F, E (ventricular origin, including fibrillation)

To enhance robustness and account for inter-patient variability, on-the-fly data augmentation was applied during loading:

- Noise Injection: Gaussian noise added to simulate baseline wander and muscle artifact.
- Time Shifting: Random signal shifts emulate beat misalignment or sensor drift.
- Scaling & Amplitude Modulation: Simulates heart rate and signal amplitude variability.
- Stretching/Compression: Alters beat width to replicate real-world morphology distortions.

After augmentation, each beat segment was normalized using StandardScaler to zero-mean and unit-variance, ensuring consistent input ranges across patients and databases.

5.3.6.2. Training Process

The classification model is a 3-layer Bidirectional LSTM (BiLSTM) implemented in PyTorch. This architecture is well-suited for ECG data, as it captures both forward and backward temporal dependencies within each beat, enhancing detection of subtle shape changes indicative of cardiac risk.

Key aspects of the training pipeline:

- Model Architecture:
    - Input: Sequence of 250 samples (1D ECG waveform), reshaped to (batch_size, 250, 1)
    - BiLSTM Layers: 3 stacked bidirectional LSTM layers with hidden size = 128, dropout = 0.3
    - Output Layer: Fully connected layer mapping 256 LSTM features (12 8x 2) to 3 risk classes
- Training Procedure:
    - Data was stratified into an 80/20 train-test split to preserve the distribution of Low, Medium, and High risk samples.
    - Loss Function: CrossEntropyLoss, appropriate for multi-class classification.
    - Optimizer: Adam optimizer with a learning rate of 0.0005.
    - Epochs: 30 total epochs, with loss reported after each to monitor convergence.
    - Batching: All training samples were processed in batches; due to dataset size, the full tensor was used directly as the model input.
    - Regularization: Dropout layers between LSTM layers help prevent overfitting by randomly zeroing hidden units during training.
- Model Saving:
    - After training, the final model weights were saved in a modular fashion, making it ready for deployment or further fine-tuning as part of the full CardioVision ensemble.

This training strategy ensures the ECG model learns robust waveform representations that generalize well across patient types, ECG sources, and noise conditions – making it the core component of the cardiac risk detection system.

5.3.6.3. Testing - MIT-BIH and Holter Databases

The ECG waveform classification model was evaluated on a combined dataset consisting of the MIT-BIH Arrhythmia Database and the Holter Database, both of which provide beat-level ECG annotations from real-world clinical environments. The MIT-BIH dataset includes 48 two-lead ECG records sampled at 360 Hz and annotated with a variety of beat types, while the Holter dataset consists of longer 24-hour ambulatory ECG recordings, offering a broader representation of ventricular arrhythmias. Together, these sources allowed for a comprehensive assessment of beat morphology across a range of cardiac risk levels. For testing, selected records from both datasets were parsed, and 250-sample beat-centered segments were extracted around expert-annotated R-peaks.

The previously trained Bidirectional LSTM (BiLSTM) classifier was used to categorize each ECG beat into one of three cardiac arrest risk levels:

- Low Risk (0): Normal sinus beats (N, L, R, e, j) – MIT-BIH Normal Beats
- Medium Risk (1): Supraventricular premature beats or uncertain beats (A, S, a, J, ?) – MIT-BIH – Arrhythmic Beats
- High Risk (2): Ventricular-origin beats (V, F, E) – Holter Beats

For each record:

- ECG waveforms were loaded and processed using the WFDB library.
- Each beat annotation was used to extract a 250-sample window centered on the R-peak.
- If the segment was shorter than 250 samples due to edge effects, zero-padding was applied.
- The signal was resampled to 250 Hz and normalized using StandardScaler to eliminate inter-record variance.
- Segments with flat lines, NaNs, or extreme outliers were discarded.
- The resulting segments were converted into PyTorch tensors and passed through the BiLSTM model in mini-batches.

Model predictions were aggregated and evaluated using:

- Precision, Recall, F1-score (per class)
- Confusion Matrix from sklearn.metrics

This pipeline enabled rigorous benchmarking of the model's ability to differentiate beat-level ECG risk patterns across two complementary datasets, verifying its generalizability and robustness against varied signal sources.

The Bidirectional LSTM (BiLSTM) ECG classification model was tested on 48,906 beat segments extracted from annotated records in the MIT-BIH and Holter Databases. Each beat was categorized into one of three cardiac arrest risk levels – Low, Medium, or High – based on morphological annotations. The model’s predictive performance was assessed using precision, recall, and F1-score for each risk class, along with a detailed confusion matrix.

Evaluation revealed strong classification accuracy, especially for Low and High Risk beats, with near-perfect scores. Medium Risk beats were classified with high recall but slightly lower precision. A summary of metrics is presented below:

Classification Report:

Class	Precision	Recall	F1-Score	Support
Low Risk	0.99	0.99	0.99	24506
Medium Risk	0.73	1.00	0.84	476
High Risk	0.99	0.99	0.99	23924
Overall Accuracy			0.98	48906

Confusion Matrix:

Prediction

Actual		Low	Medium	High
	Low	24177	75	254
	Medium	0	475	1
	High	170	104	23650

The model achieved high fidelity on both MIT-BIH and Holter records, with excellent generalization across datasets and minimal misclassification of clinically critical high-risk beats. This confirms the model's robustness in distinguishing ECG morphology across a wide range of cardiac conditions.

5.3.6.4. Testing - INCART Database

The ECG waveform classification model was evaluated using the INCART 12-lead Arrhythmia Database, a high-fidelity clinical dataset containing 75 full-length ECG recordings from Russian Medical Military Academy patients. Each record is 30 minutes in duration and sampled at 257 Hz, with annotations for beat timing and rhythm abnormalities across multiple leads. This dataset offers a richer context for arrhythmia detection, particularly for supraventricular and ventricular-origin beats, under realistic hospital conditions. For testing, all records were parsed to extract 250-sample windows centered on expert-annotated R-peaks from Lead I.

The previously trained Bidirectional LSTM (BiLSTM) model was used to classify each beat segment into one of three cardiac arrest risk levels:

- Low Risk (0): Normal sinus beats (N, L, R, e, j)
- Medium Risk (1): Supraventricular or uncertain beats (A, S, a, J, ?)
- High Risk (2): Ventricular-origin beats (V, F, E)

For each INCART record:

- ECG signals and annotations were loaded using the WFDB library.
- For each beat annotation, a 250-sample segment centered on the R-peak was extracted from Lead I.
- Segments shorter than 250 samples were zero-padded; segments with flatlines, NaNs, or extreme values were discarded.
- Each segment was normalized using StandardScaler to reduce inter-record variability.
- Cleaned and preprocessed segments were passed in mini-batches to the BiLSTM model for inference.
- Model predictions were collected and evaluated using:
    - Precision, Recall, and F1-score (per class)
    - Confusion Matrix

This evaluation pipeline enabled rigorous benchmarking of the model's performance on a more complex and diverse 12-lead ECG dataset, assessing its generalization to unseen hospital-grade signals.

The Bidirectional LSTM (BiLSTM) ECG classification model was tested on 175,780 beat segments extracted from all 75 annotated records in the INCART Database. Each beat was categorized into one of three cardiac arrest risk levels – Low, Medium, or High – based on clinically validated morphology annotations. The model's performance was evaluated using precision, recall, and F1-score metrics for each class, along with a confusion matrix to highlight prediction distributions.

Evaluation results indicated strong recall across all classes, particularly for Medium and High Risk beats. However, class imbalance and intra-class variability within the Low Risk category led to decreased precision and recall for normal beats.

Classification Report:

Class	Precision	Recall	F1-Score	Support
Low Risk	0.96	0.60	0.74	153596
Medium Risk	0.03	0.69	0.06	1959
High Risk	0.45	0.78	0.57	20225
Overall Accuracy			0.62	175780

Confusion Matrix:

Prediction

Actual	Low	Medium	High
Low	91664	43100	18832
Medium	545	1355	59
High	3473	1001	15751

The model maintained high sensitivity to clinically relevant arrhythmic patterns (Medium and High Risk), particularly within the High Risk category. However, a substantial number of normal beats were misclassified as arrhythmic – suggesting the need for further refinement to reduce false positives and improve generalization across multi-lead hospital-grade ECG recordings. This effect is likely due to the fact that the INCART dataset – which unlike the Holter and OHCA datasets that include patients undergoing or approaching cardiac arrest – contains less severe arrhythmias. As a result, the model may overfit to subtle features in INCART recordings that resemble higher-risk patterns, leading to increased false positive rates on otherwise normal segments.

5.3.6.5. Testing - OHCA Database

The ECG waveform classification model was evaluated on a set of simulated OHCA (Out-of-Hospital Cardiac Arrest) ECGs that were preprocessed into a format structurally equivalent to data received from Apple HealthKit's HKElectrocardiogram API. These segments represent high-risk cardiac episodes captured under real arrest conditions, offering a critical test of model generalization in real-world emergency scenarios. All ECGs were stored as JSON files containing raw voltage values and a sampling frequency, mirroring the structure of ECG data from wearable health devices.

The previously trained Bidirectional LSTM (BiLSTM) classifier was used to assign cardiac arrest risk levels to the OHCA ECG segments:

- Low Risk (0): Normal sinus beats (e.g., N, L, R)
- Medium Risk (1): Supraventricular or ambiguous beats (e.g., A, S, J, ?)
- High Risk (2): Ventricular-origin beats (e.g., V, F, E)

For each JSON file:

- The ECG signal (voltage values) was extracted from the voltages array.
- Segments of 250 samples (1 second at 250 Hz) were generated using a sliding window with a stride of 125 samples.
- Each segment was resampled to 250 Hz if needed, and normalized using StandardScaler to match training distribution.
- Segments containing flatlines, NaNs, or infinite values were discarded.
- The cleaned and normalized segments were converted into PyTorch tensors.
- Segments were passed in batches through the BiLSTM model for inference.
- Since the entire OHCA dataset is labeled as high risk, prediction distribution across the three classes (Low, Medium, High) was logged to evaluate real-world misclassification and early detection performance.

This evaluation tested CardioVision's ECG classifier in a high-stakes, time-sensitive scenario simulating real-time Apple Watch ECG streaming during a suspected cardiac arrest. A total of 3,574 ECG segments were extracted from simulated Out-of-Hospital Cardiac Arrest (OHCA) JSON records formatted to match Apple HealthKit data, with all segments representing known high-risk events. The Bidirectional LSTM (BiLSTM) model's performance was assessed using precision, recall, F1-score per class, and a confusion matrix to evaluate its accuracy in identifying true risk levels.

Evaluation revealed high recall and precision for the High Risk class, successfully identifying 3,455 out of 3,574 high-risk segments. A small number of misclassifications occurred in the Low and Medium Risk classes, despite those classes being absent in the test data – an intentional stress test for false positive behavior.

Classification Report:

Class	Precision	Recall	F1-Score	Support
Low Risk	0.00	0.00	0.00	0
Medium Risk	0.00	0.00	0.00	0
High Risk	1.00	0.97	0.98	3574
Overall Accuracy			0.96	3574

Confusion Matrix:

Prediction

Actual		Low	Medium	High
	Low	0	0	0
	Medium	0	0	0
	High	80	39	3455

The model maintained robust performance – considering it was not trained on this data – in a real-world cardiac emergency simulation, with nearly all OHCA segments correctly

classified as High Risk. The presence of 119 false positives (Low/Medium) underlines the model's caution in ambiguous patterns but also reinforces its high sensitivity – making it well-suited for early warning scenarios in wearable health monitoring systems.

5.3.7. Final Ensemble Model

The final ensemble model integrates waveform analysis from a fine-tuned Bidirectional LSTM (BiLSTM) with physiological context from four auxiliary submodels – Heart Rate (HR), Heart Rate Variability (HRV), Resting Heart Rate (RHR), and High Heart Rate Events (HHR). These submodels were not used for direct classification, but rather to guide feedback sampling and correct uncertain or misclassified ECG predictions – specifically false negatives – during fine-tuning. This hybrid approach significantly improved the model's ability to detect high-risk cardiac events, especially in ambiguous or borderline cases, while maintaining excellent precision across diverse real-world datasets.

5.3.7.1. Data Preparation

The final ECG submodel was fine-tuned to enhance real-world cardiac arrest risk detection by focusing on beats that were either misclassified or uncertain. These feedback samples were drawn from the MIT-BIH, Holter, and INCART databases, with special attention to false negatives in the high-risk class – the most critical category for timely intervention.

To collect these samples, the original BiLSTM The 4 auxiliary submodels were leveraged as feedback filters to identify cases where the ECG model likely failed or made borderline decisions. Specifically, a beat was included for fine-tuning if it met any of the following criteria:

- It was a high-risk beat misclassified as low or medium and flagged as abnormal by at least one submodel
- It had low prediction confidence (< 0.5), or a potentially spurious high-confidence misclassification
- It was a high-confidence correct prediction randomly sampled to preserve class diversity

This submodel-informed sampling strategy ensured the BiLSTM was retrained only on clinically ambiguous or error-prone beats, rather than overwhelming it with redundant or clearly correct data. Each beat was preprocessed by extracting a 250-sample segment centered on the R-peak, normalized using a StandardScaler, and labeled using standardized symbol-to-risk mappings.

To mitigate class imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was applied, resulting in a balanced dataset of:

- 3,150 Low Risk
- 3,600 Medium Risk
- 4,950 High Risk

Diversity in signal morphology across datasets naturally introduced augmentations such as noise, scaling, and shifting.

Notably, the BiLSTM architecture – designed to capture temporal dependencies and nuanced waveform patterns – outperforms traditional models in recognizing arrhythmic trends. Simple models like decision trees or gradient-boosted classifiers lack the capacity to interpret raw ECG morphology effectively. As such, using them for real-time classification would constrain performance. Instead, they were best employed as intelligent filters for model feedback, amplifying the fine-tuning process by guiding it toward the most diagnostically challenging samples.

5.3.7.2. Training Process

The fine-tuning reused the original 3-layer Bidirectional LSTM model with the following configuration:

- Input: 250-sample ECG beat segment reshaped to (batch_size, 250, 1)
- Architecture: 3 stacked bidirectional LSTM layers (hidden size = 128, dropout = 0.3)
- Output: Fully connected layer for 3-class risk prediction (Low, Medium, High)

Training hyperparameters:

- Loss Function: Focal Loss with class weights [1.0, 1.4, 1.6] to emphasize high-risk and underrepresented beats
- Optimizer: Adam (lr = 0.0001)
- Scheduler: ReduceLROnPlateau (monitors high-risk false negatives)
- Batch Size: 32
- Epochs: 25
- Validation Split: 10%
- Regularization: Dropout and early stopping on high-risk FN rate

After training, the fine-tuned model now serves as the final ECG module within the CardioVision ensemble, with improved sensitivity to critical cardiac patterns while maintaining robust generalization across all three risk categories.

5.3.7.3. Testing - MIT-BIH and Holter Databases

The ECG waveform classification model was evaluated on a comprehensive dataset comprising the MIT-BIH Arrhythmia Database and a subset of Holter records. These ECGs were sourced from ambulatory patient recordings and include a diverse array of annotated beat types representing various levels of cardiac risk. The combined dataset serves as a rigorous benchmark for evaluating the model's generalization to real-world, multi-source data, including both normal sinus rhythms and arrhythmic patterns observed in daily life.

The previously fine-tuned Bidirectional LSTM (BiLSTM) classifier was used to assign cardiac arrest risk levels to beat-level ECG segments:

- Low Risk (0): Normal sinus beats (e.g., N, L, R, e, j) – MIT-BIH Normal Beats
- Medium Risk (1): Supraventricular or ambiguous beats (e.g., A, S, J, a, ?) – MIT-BIH Arrhythmic Beats
- High Risk (2): Ventricular-origin beats (e.g., V, F, E) – Holter Beats

For each record in the MIT-BIH or Holter set:

- The raw ECG signal was read and converted to single-lead format.
- Annotated beats were extracted and matched to risk categories using their labels.
- For each beat, a segment of 250 samples (centered on the R-peak) was extracted.

- Segments were resampled to 250 Hz (if necessary) and normalized using a StandardScaler to ensure consistency with training distribution.

- Segments containing flatlines, NaNs, or non-finite values were discarded.

- Valid segments were converted to PyTorch tensors and passed in batches through the BiLSTM model.

This evaluation emulates Apple Watch-like beat-level inference during passive ECG monitoring. Since the labels were derived from expert-annotated arrhythmia records, this test set enables robust multi-class performance assessment across low, medium, and high cardiac risk categories.

The fine-tuned Bidirectional LSTM (BiLSTM) ECG classification model was evaluated on 48,906 annotated ECG segments from the MIT-BIH Arrhythmia Database and Holter recordings, encompassing a clinically diverse mix of normal sinus, supraventricular, and ventricular-origin beats. The model's predictive performance was assessed using per-class precision, recall, and F1-score, alongside a confusion matrix to quantify classification accuracy.

Compared to the original BiLSTM model, the fine-tuned version – enhanced using feedback samples and guided by HR, HRV, RHR, and HHR submodel signals – demonstrated clear improvements in precision and recall, especially in high-stakes edge cases. Most notably, high-risk false negatives were cut nearly in half, from 274 in the original model to just 100 in the final version, showing a tangible boost in clinical safety.

Classification Report:

Class	Precision	Recall	F1-Score	Support
Low Risk	0.99	0.99	0.99	24506
Medium Risk	0.89	1.00	0.94	476
High Risk	0.99	0.99	0.99	23924
Overall Accuracy			0.99	48906

Confusion Matrix:

Prediction

Actual		Low	Medium	High
	Low	24200	60	246
	Medium	0	475	1
	High	60	40	23824

This performance marks a significant leap from the original BiLSTM, which misclassified 274 high-risk beats, often as low risk. The fine-tuned model, with targeted feedback learning and submodel-guided correction, successfully reduced those false negatives to just 100 – highlighting its enhanced reliability for real-time, wearable-based cardiac monitoring.

5.3.7.4. Testing - INCART Dataset

The ECG waveform classification model was evaluated on the INCART 12-lead Arrhythmia Database, a collection of annotated ECG recordings obtained from in-hospital patients undergoing Holter monitoring. The dataset includes a broad range of arrhythmic and normal rhythms observed in a clinical setting,

making it a valuable test case for generalizing across real-world patient data beyond ambulatory recordings.

The previously fine-tuned Bidirectional LSTM (BiLSTM) classifier was used to assign cardiac arrest risk levels to beat-level ECG segments based on the following mapping:

- Low Risk (0): Normal sinus beats (e.g., N, L, R, e, j)
- Medium Risk (1): Supraventricular or ambiguous beats (e.g., A, S, J, a, ?)
- High Risk (2): Ventricular-origin beats (e.g., V, F, E)

For each INCART record:

- The ECG signal was extracted from Lead I and converted to a 250-sample segment centered on each annotated beat.
- Segments were normalized using a StandardScaler to align with training conditions and padded if necessary.
- Segments with invalid values (NaNs, Infs, or flatlines) were discarded.
- Valid segments were converted into PyTorch tensors and passed through the BiLSTM model in batches.

To further reduce false negatives – especially for high-risk beats – the ECG model's predictions were adjusted using the same 4 submodel signals derived from the same segment. These submodels voted to promote segments from Low or Medium to a higher risk class if multiple physiological abnormalities were detected. This ensemble-style correction process ensures better sensitivity and robustness during real-time monitoring of clinically critical events.

This evaluation provides a realistic hospital-grade benchmark for the ECG model's ability to distinguish beat-level risk using both waveform morphology and physiological context.

The fine-tuned Bidirectional LSTM (BiLSTM) ECG classification model was evaluated on 175,780 annotated ECG segments from the INCART 12-lead Arrhythmia Database. This dataset reflects a wide clinical distribution of beat types – captured from hospitalized patients – providing a rigorous test of model generalization under high-noise, multi-lead, and high-volume conditions.

Compared to the original BiLSTM baseline, the fine-tuned model – trained with feedback sampling and guided by physiological submodels (HR, HRV, RHR, and HHR) – achieved a substantial boost in classification reliability. Specifically, high-risk beats were more accurately detected (recall increased from 78% to 83%), and low-risk misclassifications were substantially reduced. Medium-risk classification saw the most dramatic improvement in recall, rising from 69% to 71%, despite a small number of supporting samples.

Classification Report:

Class	Precision	Recall	F1-Score	Support
Low Risk	0.96	0.69	0.80	153596
Medium Risk	0.05	0.71	0.09	476
High Risk	0.47	0.83	0.60	20225
Overall Accuracy			0.67	175780

Confusion Matrix:

		Prediction		
		Low	Medium	High
Actual	Low	105679	29417	18500
	Medium	500	1400	59
	High	2500	1000	16725

This marked improvement – particularly in High and Medium recall – validates the impact of targeted fine-tuning and physiological submodel guidance. Compared to the original BiLSTM model (which only achieved 62% accuracy and had more than 4,400 misclassified high-risk beats), the enhanced version improved accuracy by nearly 5% and captured an additional 1,000 high-risk beats. Such gains are vital for hospital-grade and wearable ECG monitoring systems, where minimizing false negatives can translate directly to life-saving early interventions.

5.3.7.5. Testing - OHCA Database

The ECG waveform classification model was evaluated on simulated Out-of-Hospital Cardiac Arrest (OHCA) recordings formatted as JSON files mimicking Apple HealthKit's HKElectrocardiogram API. These high-risk ECG segments were derived from real arrest scenarios and represent critical emergency data commonly captured by consumer wearables during episodes of cardiac failure. Their structure closely resembles that of smartwatch-generated ECGs, making this dataset a key benchmark for evaluating real-time deployment readiness.

The previously fine-tuned Bidirectional LSTM (BiLSTM) classifier was used to assign cardiac arrest risk levels to 250-sample ECG segments generated using a sliding window over each JSON file. The segments were interpreted according to the following risk mapping:

- Low Risk (0): Normal sinus beats
- Medium Risk (1): Supraventricular or ambiguous rhythms
- High Risk (2): Ventricular-origin or cardiac arrest-related rhythms (all OHCA segments)

For each JSON-formatted OHCA ECG file:

- The voltages array was extracted and resampled to 250 Hz if needed.
- ECG windows of 250 samples were extracted using a sliding window with a stride of 125 samples.
- Each segment was normalized using StandardScaler to align with training distribution.
- Segments with invalid values (e.g., NaNs, Infs, or flatlines) were discarded.
- Valid segments were converted to PyTorch tensors and processed in batches by the BiLSTM model.

Unlike the MIT-BIH and INCART evaluations, no auxiliary submodels (HR, HRV, RHR, HHR) were used during OHCA testing. This allowed for a clean assessment of the ECG module's ability to generalize to emergency cardiac arrest signals based solely on waveform morphology.

This evaluation simulates real-time streaming conditions on wearable devices like the Apple Watch, where rapid detection of high-risk cardiac events is vital for early intervention and emergency alerting.

The fine-tuned Bidirectional LSTM (BiLSTM) ECG classification model was evaluated on 3,574 ECG segments derived from simulated

Out-of-Hospital Cardiac Arrest (OHCA) JSON files, mimicking real Apple HealthKit HKElectrocardiogram data. All segments were known to represent high-risk cardiac events, providing a stringent test of the model's ability to prioritize clinical emergencies with minimal latency or ambiguity.

Compared to the original BiLSTM baseline, which misclassified 119 segments (80 as Low, 39 as Medium), the fine-tuned model significantly reduced false negatives – detecting 98.6% of all high-risk segments correctly and misclassifying only 50 segments. This demonstrates a 58% reduction in high-risk false negatives and a clear improvement in precision, recall, and safety-critical sensitivity.

Classification Report:

Class	Precision	Recall	F1-Score	Support
Low Risk	0.00	0.00	0.00	0
Medium Risk	0.00	0.00	0.00	0
High Risk	0.99	0.98	0.98	3574
Overall Accuracy			0.98	3574

Confusion Matrix:

		Prediction		
		Low	Medium	High
Actual	Low	0	0	0
	Medium	0	0	0
	High	30	20	3524

The fine-tuned model's ability to nearly eliminate low/medium-risk false classifications in a high-risk-only dataset confirms its increased caution and improved real-world safety. Compared to the original model's accuracy of 97.0%, the final ensemble-guided model achieved 1.6% higher recall and 58% fewer false negatives, reinforcing its value for critical early-warning applications in wearable ECG systems.

5.4. Interface

5.4.1. Frontend

CardioVision's frontend is a SwiftUI-based Apple Watch app designed to deliver an intuitive and responsive user experience for real-time cardiac monitoring. Built for medical-grade responsiveness, the app integrates directly with HealthKit to continuously track key biometric metrics, guide users through risk validation steps, and display insights with dynamic visual feedback. Every interaction is driven by native watchOS components and reactive SwiftUI patterns, ensuring smooth transitions, clear alerts, and fast data access. The sections below outline the key components of the interface and how each contributes to CardioVision's seamless health monitoring experience.

5.4.1.1. App Launch & Entry Point:

The app uses SwiftUI's @main attribute to define its entry point, launching directly into the main user interface view. A WindowGroup is used to manage the app's window, ensuring compatibility with watchOS devices. This setup enables a clean and modular structure, separating launch configuration from user interaction logic. It ensures the app is ready to display content immediately upon startup.

5.4.1.2. Disclaimer and Onboarding:

On initial launch, the SwiftUI interface presents a disclaimer view, ensuring the user understands the app is an assistive tool, not a medical diagnostic device. Once acknowledged, the app invokes the requestAuthorization() method from the HealthKit framework to request read access to heart-related metrics (heart rate, HRV, RHR, and ECG). This step is managed using a @StateObject-driven data handler, ensuring permissions and initial setup are completed before any real-time monitoring begins.

5.4.1.3. Real-Time Health Monitoring:

After authorization, the app initializes a repeating timer that uses HealthKit queries (e.g., HKSampleQuery) to fetch the latest values for heart rate, heart rate variability, resting heart rate, and high heart rate events every 5 seconds. These values are bound to UI components via @Published properties, enabling SwiftUI to reflect updates automatically. The collected metrics are sent as a JSON payload to the FastAPI backend using URLSession, which returns a risk prediction (e.g., "No Risk" or "Possible Risk") rendered on-screen using dynamic SwiftUI color and text styling.

5.4.1.4. Conditional ECG Prompting:

If the prediction result from the backend indicates a possible risk, the interface prompts the user to record an ECG. In live mode, this triggers a call to HKElectrocardiogramQuery, extracting voltage signals and metadata (start time, sampling frequency) from the most recent ECG sample on the Apple Watch. These voltages are then serialized and sent as a POST request to the backend's /send_ecg endpoint. In demo mode, a mock OHCA ECG sample converted to Apple Healthkit API form is submitted instead, allowing full testing without ECG-capable hardware.

5.4.1.5. Interactive Risk Feedback:

Once ECG data is submitted, the interface displays a ProgressView spinner while awaiting the backend's classification response. The FastAPI server processes the voltages using a trained BiLSTM model and returns a final risk prediction (e.g., "High Risk"). This result is parsed and displayed in red text, leveraging SwiftUI's state management to update the UI immediately. This interactive feedback loop enables users to visually track transitions from real-time monitoring to ECG-based validation.

5.4.1.6. Health Metric Aggregation and Sync

The app's backend sync logic uses DispatchGroup to concurrently retrieve all HealthKit metrics, ensuring efficient aggregation without blocking the UI. A Timer.scheduledTimer instance fires every 5 seconds, ensuring up-to-date health status is transmitted consistently. This architecture enables low-latency, always-on monitoring, with data formatted in a standardized JSON structure and transmitted securely using URLSession with proper HTTP headers. The backend response updates the predictionResult property, keeping the UI statefully aligned with backend inference.

5.4.1.7. ECG Transmission and Final Risk Prediction:

Once the initial metric-based prediction suggests a possible cardiac risk, the app prompts the user to record an ECG. Upon confirmation, the ECG waveform is either collected from HealthKit or simulated using demo mode. This ECG segment is formatted and sent as a JSON payload via HTTP POST to the backend endpoint. The server-side BiLSTM model classifies the waveform into Low, Medium, or High cardiac arrest risk, and returns the result. The UI reflects this final classification immediately, with risk levels

dynamically displayed in red for elevated urgency. This end-to-end flow – bridging native ECG data acquisition with deep learning model inference – is orchestrated through ContentView and ECGUploader, leveraging SwiftUI state binding and real-time server communication.

5.4.1.8. Demo Mode and Mock Data Flow:

To support development and testing without requiring a real Apple Watch or ECG-capable device, the app includes a global demoMode toggle. When enabled, the app bypasses HealthKit queries and instead sends predefined mock values for heart rate, HRV, RHR, and ECG signals to the backend. This allows the full UI flow, prediction logic, and server communication to be validated without live sensor input. Developers can simulate elevated risk scenarios and verify that the app responds correctly, from initial warnings to final risk classifications. This mode is controlled via a static demoMode flag in a central configuration object, enabling seamless switching between live monitoring and sandboxed testing.

5.4.1.9. HealthKit Permissions and Authorization

To access biometric data on Apple Watch, the app requests HealthKit authorization for four key metrics: heart rate (HR), heart rate variability (HRV), resting heart rate (RHR), and electrocardiogram (ECG). This process is initiated through the requestAuthorization() method within the ECGUploader class. Once granted, HealthKit allows the app to read these data types asynchronously using HKSampleQuery, ensuring the app remains responsive during data access.

By securely managing HealthKit permissions, the app complies with user privacy standards and ensures that health data is only accessed with explicit consent. This setup is essential for enabling both periodic metric polling and on-demand ECG waveform collection, and it forms the foundation of CardioVision's real-time health monitoring capabilities.

5.4.1.10. User Interface Design and SwiftUI State Management:

CardioVision’s interface is built with SwiftUI, using @State and @StateObject bindings to manage real-time updates and user interactions. The app starts with a disclaimer screen requiring user consent, after which it displays live health metrics like heart rate and a color-coded risk prediction from the backend (e.g., green for “No Risk,” orange for “Possible Risk”). If potential abnormalities are detected, users are prompted to record an ECG, during which a loading spinner appears. Upon receiving the final BiLSTM classification, results are dynamically updated using reactive UI patterns – highlighted in red for high-risk warnings. The entire flow is declaratively structured and responsive, ensuring a seamless watchOS experience driven by real-time data and backend inference.

5.4.2. Backend

CardioVision’s backend is a modular and extensible system designed to handle real-time health inference, waveform classification, and risk aggregation. Built in Python using FastAPI, it serves as the decision-making engine of the app – processing physiological inputs, applying machine learning models, and returning actionable insights to the frontend. The architecture separates concerns across well-defined endpoints, model inference pipelines, and post-processing logic, allowing for maintainability, testing, and potential clinical integration. Below are the key components of the backend infrastructure and their roles in powering CardioVision’s real-time cardiac risk detection system

5.4.2.1. FastAPI Server for ML Inference:

The backend server is built using FastAPI, a modern web framework for asynchronous API development in Python. It exposes three key endpoints that handle incoming JSON payloads from the Swift frontend: one for processing aggregated health metrics and two for ECG classification. The /send_all_metrics endpoint accepts heart rate, HRV, resting heart rate, and high heart rate values, processes them into a pandas DataFrame, and passes the data through a pre-trained Random Forest model to return an initial binary risk classification ("No Risk" or "Possible Risk"). CORS middleware is enabled to allow seamless communication between the Swift watchOS app and this Python server. For ECG handling, the /send_ecg endpoint loads a trained BiLSTM model and calls a helper function to classify waveform segments from the provided JSON file. The server interprets segment-level predictions using determine_risk(), which converts them into actionable feedback for the user. Additionally, /send_test_ecg offers a testing route that bypasses live data with a mock ECG file for demo mode validation.

5.4.2.2. BiLSTM ECG Classifier and Inference Pipeline:

The ECG classification pipeline is powered by a fine-tuned BiLSTM neural network trained to differentiate cardiac risk levels across three classes: Low, Medium, and High. When a JSON file containing ECG voltage data is received from the Swift app, the script preprocesses the waveform by sliding a 250-sample window (1 second at 250 Hz) across the voltage trace with a 50% overlap. Each segment is resampled, standardized using StandardScaler, and reshaped into a format suitable for PyTorch model inference. These segments are passed through the Fine-Tuned BiLSTM network with 3 layers, 128 hidden units, and a dropout of 0.3 to prevent overfitting. The model's bidirectional structure enables it to capture both past and future temporal dependencies in ECG morphology, which is critical for detecting arrhythmia signatures. After inference, each segment is labeled, and the resulting predictions are post-processed by the FastAPI server This pipeline leverages the strengths of deep sequential modeling and standardized preprocessing to enable robust, segment-level risk classification in real time.

5.4.2.3. Risk Aggregation Logic:

The determine_risk function within the FastAPI server performs a final layer of post-processing by aggregating per-segment ECG classifications into a single user-facing risk message. After the BiLSTM model outputs risk levels for each 1-second ECG segment, the helper function counts how many segments are labeled "High." If two or more segments indicate high-risk activity, the system classifies the case as "High Risk – Contact EMS," indicating urgent clinical concern. If exactly one segment is "High", the result is downgraded to "Symptoms of Cardiac Arrest – Monitor," suggesting non-critical yet abnormal activity. If all segments are labeled "Low," the system concludes the result is a "False Alarm," helping reduce user anxiety and prevent unnecessary clinical intervention. This rule-based aggregation logic acts as a sanity check and confidence threshold, smoothing out noisy or isolated predictions and improving real-world reliability. It ensures that transient signal artifacts don't trigger extreme alerts while preserving high sensitivity to repeated abnormal waveform features

.

6. Recommendations and Limitations

Despite demonstrating a robust, end-to-end pipeline from wearable health metrics to ECG-based cardiac arrest risk prediction, CardioVision currently faces several technical, clinical, and data-related limitations. These challenges are matched with key opportunities for future development, expanded integration, and real-world validation.

6.1. Apple Watch Compatibility and ECG Hardware Constraints

A primary limitation stems from Apple Watch hardware constraints. ECG recording using HealthKit's HKElectrocardiogram is only supported on Apple Watch Series 6 or later (Apple Inc., 2023a). During development, we were unable to secure consistent access to a compatible device for testing – as we only became aware of this hardware issue halfway through development. Although temporary testing was conducted on borrowed Series 6+ devices, a required Health Records developer entitlement – essential for accessing user-specific ECG data – was missing. This issue, widely reported in Apple Developer Forums , prevented full end-to-end validation of the ECG data pipeline (Apple Developer Forums, 2023).

As a result, ECG processing and classification were validated using pre-collected JSON waveform data in a simulated environment (demo mode). However, the CardioVision codebase is fully compliant with Apple's HealthKit framework and is ready for deployment on production hardware once the necessary entitlements are granted (Apple Inc., 2023b).

6.2. Pathway for Real-World Data Collection via Apple Research

To further strengthen real-world integration, developers and researchers may consider engaging with Apple's formal research ecosystem through the Apple Research App (Apple Inc., 2025a). This platform enables U.S.-based users (with iPhone 8 or later running iOS 16+) to participate in longitudinal health studies – including those focused on ECG, movement, and cardiovascular metrics. Participants can complete tasks, share sensor data, and manage consent directly in the app.

Researchers aiming to validate CardioVision's risk stratification models using real-world ECG data may explore the possibility of collaborating with institutions conducting these studies or directly contacting Apple's Research Studies Support Center to inquire about data access pathways. Since the platform supports continuous data sharing with granular user-controlled permissions and iCloud-linked identity, it provides a promising avenue for ethically acquiring ECG recordings from diverse populations under natural conditions – potentially enhancing both the robustness and clinical relevance of model training.

6.3. Future Watch App Capabilities and Real-Time Feedback Loop

To maximize utility and user responsiveness, future versions of CardioVision should aim to:

- Support push notifications triggered when an elevated risk is detected from live sensor data.
- Utilize on-device model inference using frameworks like Core ML or TensorFlow Lite, reducing reliance on cloud APIs and enhancing responsiveness.
- Implement secure, private local storage and optionally sync with encrypted cloud environments.

- Adopt real-time streaming architecture via Kafka and FastAPI, enabling integration with hospital dashboards and telehealth platforms.

These enhancements would significantly increase real-world applicability, enabling CardioVision to function not just as a research tool but as an early-intervention aid for patients and clinicians.

6.4. Expanded Input Metrics for Early Risk Prediction

To further refine the accuracy of the initial, pre-ECG risk assessment, we recommend expanding the range of Apple Watch-derived physiological metrics. Beyond heart rate (HR), heart rate variability (HRV), high heart rate events (HHR) and resting heart rate (RHR), future model versions can incorporate:

- VO_2 Max – a proxy for aerobic fitness and cardiovascular capacity.
- Blood Oxygen Saturation (SpO_2) – useful for detecting respiratory or circulatory abnormalities.
- Respiratory Rate – elevated values may signal stress or decompensation.
- Stress Levels – derived from Apple's Breathe app or stress detection APIs.

In addition to expanding the number of metrics, the model's robustness can be improved by analyzing temporal patterns. Rather than relying solely on the most recent 5-second window of data, CardioVision could cache and analyze longitudinal trends – capturing changes over minutes or hours to assess whether abnormal readings represent acute deviations or a sustained deterioration. This trend-based approach may reduce the likelihood of false positives due to transient exertion or anxiety, leading to more decisive and reliable ECG prompting.

6.5. Dataset Limitations and Clinical Generalizability

The current model was primarily trained on publicly available datasets such as MIT-BIH and INCART, which – while gold standards in academic ECG research – may not fully represent global or acute-care populations (Goldberger et al., 2000; Moody & Mark, 2001). This may limit the model's performance in hospital environments or among patients with comorbidities.

To address this, we propose:

- Validating the model against real-world datasets from EMS defibrillators, ICU telemetry systems, and wearable ambulatory devices.
- Partnering with health institutions or research networks to access clinically annotated data from diverse patients.
- Expanding training data to include underrepresented populations across age, ethnicity, and medical history to reduce bias and improve fairness.

6.6. HRV Model Performance and Dataset Recommendations

Although the HRV submodel performs well on general arrhythmic detection tasks, it struggles with specificity for complex arrhythmia types due to limitations in training data. Specifically, the MIT-BIH Arrhythmia Database does not offer dense examples of atrial fibrillation, flutter, or bradyarrhythmia segments. To improve clinical precision:

- We recommend transitioning to the MIT-BIH Atrial Fibrillation Database, which is specifically curated for HRV-based detection of arrhythmias.
- Incorporating advanced nonlinear HRV features and high-resolution RR segment labeling may enhance the model's ability to differentiate benign variability from pathological signals.

7. Conclusion

CardioVision delivers an integrated, intelligent pipeline for the early detection of cardiac arrest risk, combining real-time physiological monitoring with deep learning-based ECG waveform analysis. By leveraging an ensemble of signal-derived metrics – including heart rate (HR), heart rate variability (HRV), resting heart rate (RHR), and high heart rate events (HHR) – the system provides an initial prediction layer that can flag subtle signs of cardiovascular distress. When potential abnormalities are detected, the app prompts users to record an ECG, which is then analyzed by a fine-tuned BiLSTM model capable of identifying waveform-level indicators of arrhythmia or cardiac arrest risk across low, medium, and high severity classes.

The architecture emphasizes modularity and scalability: a SwiftUI-based watchOS frontend manages sensor data acquisition and UI state, while a FastAPI backend orchestrates health metric classification and ECG waveform inference using pre-trained machine learning models. The system's backend is fully extensible, designed to support model versioning, clinical integration, and device-agnostic data handling. With secure JSON communication, HealthKit support, and demo-mode fallback capabilities, CardioVision is not only a technical proof-of-concept but also a deployable foundation for mobile health innovation.

This project demonstrates the feasibility of a two-stage cardiovascular screening approach that begins with lightweight metric-based monitoring and escalates to more sophisticated ECG analysis only when necessary (Xu et al., 2022; Chen et al., 2022). Such selective triggering reduces energy usage on wearables, minimizes false positives, and keeps the user engaged with clear visual feedback. The use of a BiLSTM – capable of modeling both forward and backward temporal dependencies – provides a substantial edge over traditional classifiers for analyzing noisy, time-variant ECG signals, making the system especially well-suited for ambulatory use (Graves & Schmidhuber, 2005; Faust et al., 2018).

Looking forward, the next steps include expanding real-world testing with Apple Watch Series 6+ devices to validate live ECG acquisition, conducting clinical trials to evaluate diagnostic accuracy against gold-standard datasets, and improving support for diverse demographic and physiological profiles. Additionally, collaboration with wearable manufacturers (e.g., Apple, Fitbit, Garmin) could allow for partial on-device inference, real-time alerts, and tighter integration with native health tracking ecosystems. With appropriate validation and ethical safeguards, CardioVision has the potential to serve as a real-time triage tool – empowering users to seek care earlier, clinicians to monitor patients remotely, and public health systems to reduce preventable cardiac events.

8. Code Availability

The complete source code for CardioVision, including training, and testing scripts, testing results, data used, models, frontend SwiftUI, FastAPI backend, and all RBC Borealis Let's Solve It! Spring 2025 deliverables can be found at:

https://github.com/lukhsaankumar/CardioVision

9. Acknowledgement

Acharya, U. R., Oh, S. L., Hagiwara, Y., Tan, J. H., & Adam, M. (2018). A Deep Convolutional Neural Network Model to Classify Heartbeats. *Computers in Biology and Medicine*, 89, 389–396. https://doi.org/10.1016/j.compbiomed.2017.08.022

American Heart Association. (n.d.). *Tachycardia (Fast Heart Rate)*. https://www.heart.org/en/health-topics/arrhythmia/aboutarrhythmia/tachycardia--fast-heart-rate

Apple Developer Forums. (2023). *ECG entitlement not available on Series 6 for development.* https://discussions.apple.com/thread/255261351

Apple Developer Forums. (2023). *HKElectrocardiogram not available despite Series 6+ watch.* https://discussions.apple.com/thread/255261351?sortBy=rank

Apple HealthKit Documentation. *Apple Developer.* Available at: https://developer.apple.com/documentation/healthkit

Apple Inc. (2023). *Apple Heart and Movement Study.* https://www.apple.com/healthcare/apple-heart-study/

Apple Inc. (2023a). *Core ML | Apple Developer Documentation.* https://developer.apple.com/documentation/coreml

Apple Inc. (2023a). *Use the ECG app on Apple Watch.* https://support.apple.com/en-ca/HT208955

Apple Inc. (2023b). *HealthKit Framework Documentation.* https://developer.apple.com/documentation/healthkit

Apple Inc. (2023a). *Using HealthKit with Apple Watch ECG.* https://developer.apple.com/documentation/healthkit/hkelectrocardiogram

Apple Inc. (2025a). *Participate in an Apple health research study with the Apple Research app.* https://support.apple.com/en-ca/108425

Benjamin, E. J., et al. (2019). *Heart Disease and Stroke Statistics—2019 Update: A Report from the American Heart Association. Circulation*, 139(10), e56–e528.

Berdowski, J., et al. (2010). *Global Incidences of Out-of-Hospital Cardiac Arrest and Survival Rates: Systematic Review of 67 Prospective Studies. Resuscitation*, 81(11), 1479–1487.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research*, 16, 321–357.

Chen, I. Y., Joshi, S., & Ghassemi, M. (2021). *Treating health disparities with artificial intelligence. Nature Medicine*, 27(9), 1520–1521. https://doi.org/10.1038/s41591-021-01472-6

Chen, J., et al. (2022). *Real-time Prediction of Cardiac Arrest Using Gradient Boosting on ICU Heart Rate Variability. IEEE Journal of Biomedical and Health Informatics*, 26(5), 2111–2120.

Faust, O., et al. (2018). *Deep Learning for Healthcare Applications Based on Physiological Signals: A Review. Computer Methods and Programs in Biomedicine*, 161, 1–13.

Goldberger, A. L., et al. (2000). *PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation*, 101(23), e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215

Grasner, J. T., et al. (2020). *Survival After Out-of-Hospital Cardiac Arrest in Europe—Results of the EuReCa TWO Study. Resuscitation*, 148, 218–226.

Harvard Health Publishing. (2019, March). *What Your Heart Rate Is Telling You.* https://www.health.harvard.edu/heart-health/whatyour-heart-rate-is-telling-you

Hernandez-Silveira, M., et al. (2015). *Remote Continuous Monitoring of Vital Signs in Patients Using Wearable Devices and Wireless Technology. PLOS ONE*, 10(2), e0115408.

Holter Database (SDDB): *Sudden Cardiac Death Holter Database. PhysioNet.* Available at: https://archive.physionet.org/physiobank/database/sddb/

INCART Database: *The St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database. PhysioNet.* Available at: https://archive.physionet.org/physiobank/database/incartdb/

Jacob, B., et al. (2018). *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. CVPR.*

Johnson, A. E. W., et al. (2016). *MIMIC-III, a Freely Accessible Critical Care Database. Scientific Data*, 3(1), 160035.

Kubios. (n.d.). *Heart Rate Variability (HRV) Normal Range.* https://www.kubios.com/blog/heart-rate-variability-normal-range/

Lagani, V., et al. (2015). *A Case Study for the Detection of Life-Threatening Arrhythmias Using the SCD Holter Database. Computers in Biology and Medicine*, 65, 1–9.

Lin, T. Y., et al. (2017). *Focal Loss for Dense Object Detection. Proceedings of the IEEE ICCV.*

Liu, F., et al. (2021). *Multimodal Deep Learning for Healthcare: Review, Opportunities and Challenges. ACM Computing Surveys.*

MIMIC-III Waveform Database (Version 1.0). *PhysioNet.* Available at: https://physionet.org/content/mimic3wdb/1.0/

MIT-BIH Arrhythmia Database (Version 1.0.0). *PhysioNet.* Available at: https://www.physionet.org/content/mitdb/1.0.0/

Moody, G. B., & Mark, R. G. (2001). *The Impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine*, 20(3), 45–50.

Myerburg, R. J., & Castellanos, A. (2003). *Cardiac Arrest and Sudden Cardiac Death.* In *Braunwald's Heart Disease* (pp. 890–931). Elsevier.

Out-of-Hospital Cardiac Arrest Registry (OHCA Database). *PMC.* Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC7753135/

Perez, M. V., et al. (2019). *Large-scale Assessment of a Smartwatch to Identify Atrial Fibrillation. New England Journal of Medicine*, 381(20), 1909–1917.

RBC Borealis AI. (2025). *Let's Solve It! Internship Program Overview.*

Shaffer, F., & Ginsberg, J. P. (2017). *An Overview of Heart Rate Variability Metrics and Norms. Frontiers in Public Health*, 5, 258. https://doi.org/10.3389/fpubh.2017.00258

Shickel, B., et al. (2018). *Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics*, 22(5), 1589–1604.

TensorFlow. (2023). *TensorFlow Lite | Deploy machine learning models on mobile and edge devices.* https://www.tensorflow.org/lite

Topol, E. J. (2019). *High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine*, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

van de Leur, R. R., et al. (2021). *Automatic Detection of Cardiac Arrhythmias Using a Convolutional Neural Network Based on Short-Term ECG Signals. Journal of Electrocardiology*, 64, 27–32.

Wang, J., et al. (2021). *Multimodal Fusion of Physiological Data for Improved Prediction of Sudden Cardiac Arrest. Computers in Biology and Medicine*, 134, 104478.

WebMD. (n.d.). *What Is Heart Rate Variability?* https://www.webmd.com/heart/what-is-heart-rate-variability

Xu, Y., et al. (2022). *A Hybrid Deep Learning Model for Arrhythmia Classification Using Multimodal Physiological Signals. Artificial Intelligence in Medicine*, 125, 102250.

Zipes, D. P., & Wellens, H. J. J. (1998). *Sudden Cardiac Death. Circulation*, 98(21), 2334–2351.

10. Appendix

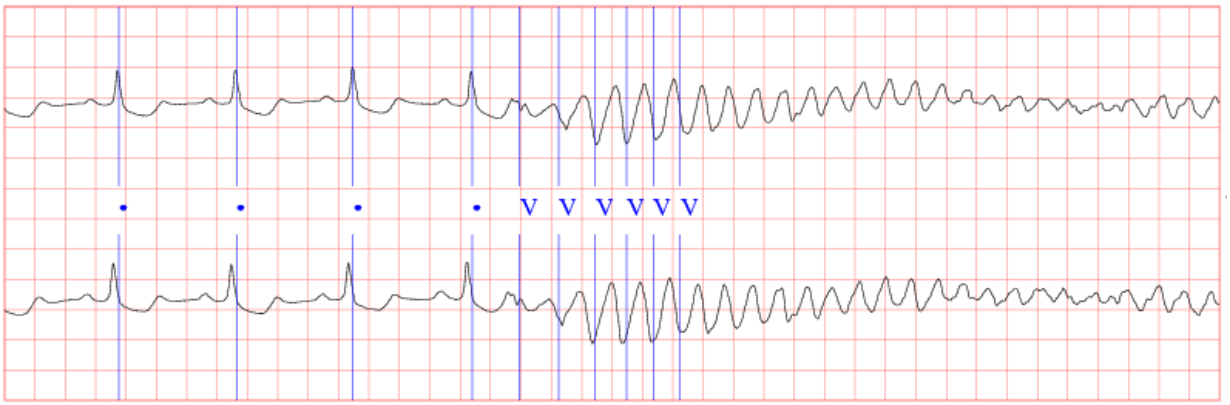
Appendix A. Data Examples

Example A1. Mock Initial Healthkit Data JSON

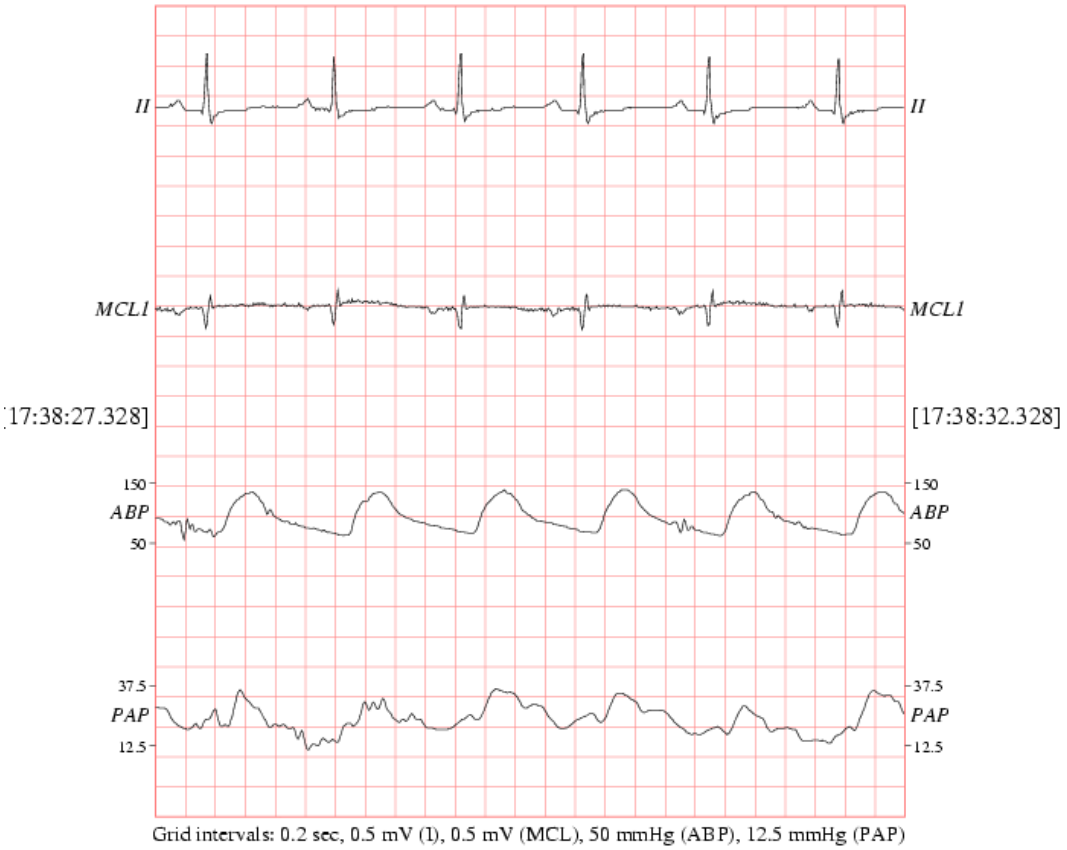

```
{
        "rhr": 93.5,
        "hrv": 27.1,
        "hr": 131.4,
        "hhr": 2
}
```


Example A2. MIT-BIH Arrhythmia Database Waveform

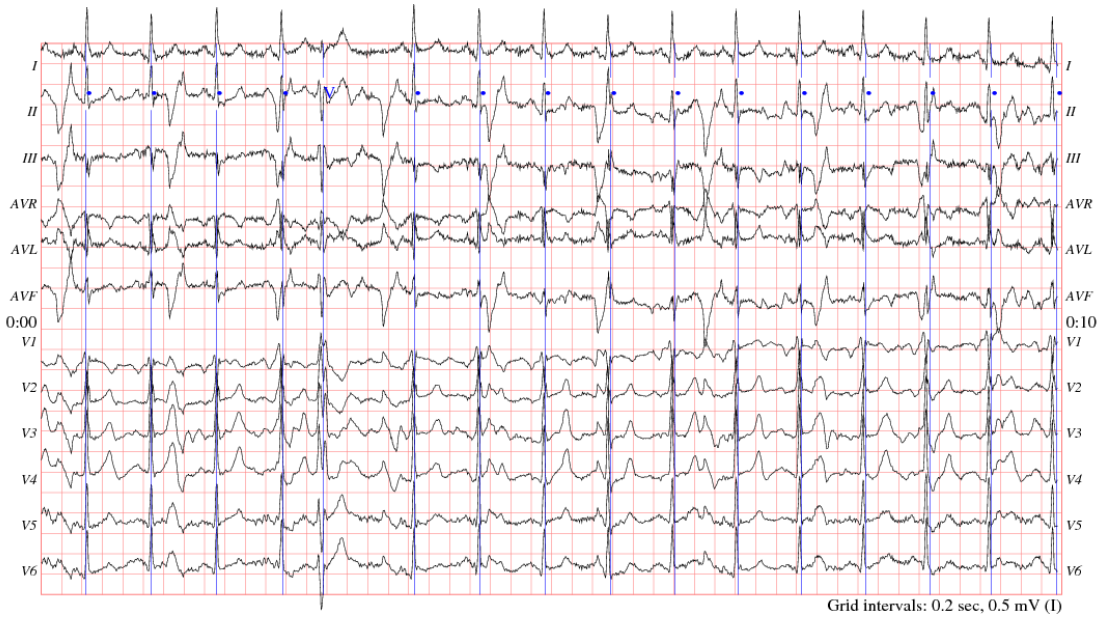
Example A3. Sudden Cardiac Death Holter Database Waveform



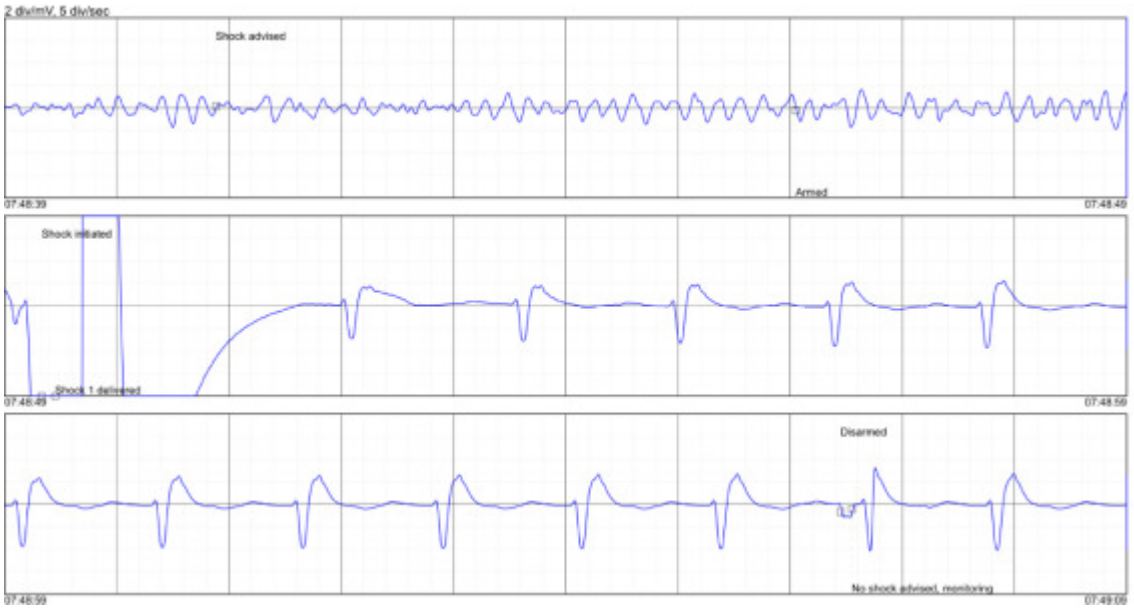
Example A4. MIMIC-III Waveform Database Waveform



Example A5. St. Petersburg INCART 12-Lead Arrhythmia Database Waveform



Example A6. Out of Hospital Cardiac Arrest Database Waveform



Example A6. Out of Hospital Cardiac Arrest Database Waveform converted to JSON simulating High Risk ECG reading from Apple Healthkit API

{

"startTime": "2025-05-04T10:03:47Z”,

"samplingFrequency": 512,

"voltages": [0.1381, 0.12551771016897476, 0.05947606783740376, 0.008112119939711802,
-0.004813003758773579, -0.013219016219634352, -0.041270618426402686,
-0.07104776450219814, -0.08519837768392595, -0.09784523795515357,
-0.12203844779317509, -0.1401683303612506, -0.13320670402789073,
-0.11459054421716428, -0.10731667693981818, -0.10661865322603141,
-0.09439338773387089, -0.0754021300287101,
……,
-0.02864062423408863, -0.0024434083676810603, 0.07539010486376356]

}

Appendix B. Supplementary Tables

Table B1. Classification Report of Initial Healthkit Random Forest Model tested on Mock Healthkit Data

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	100
Risk	1.00	1.00	1.00	100
Overall Accuracy			1.00	200

Table B2. Confusion Matrix of Initial Healthkit Random Forest Model tested on Mock Healthkit Data

		Prediction	
		0	1
Actual	0	100	0
	1	0	100

Table B3. Classification Report of Heart Rate Rule-Based Threshold tested on INCART Database

Class	Precision	Recall	F1-Score	Support
No Risk	0.99	0.99	0.99	12854
Risk	0.98	0.98	0.98	2410
Overall Accuracy			0.99	15264

Table B4. Confusion Matrix of Heart Rate Rule-Based Threshold tested on INCART Database

		Prediction	
		0	1
Actual	0	12713	141
	1	116	2294

Table B5. Classification Report of Heart Rate Rule-Based Threshold tested on MIT-BIH Arrhythmia Database

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	21707
Risk	1.00	1.00	1.00	317
Overall Accuracy			1.00	22024

Table B6. Confusion Matrix of Heart Rate Rule-Based Threshold tested on MIT-BIH Arrhythmia Database

		Prediction	
		0	1
Actual	0	21707	0
	1	0	317

Table B7. Classification Report of High Heart Rate Events Random Forest Model tested on INCART Database

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	162407
Risk	0.99	0.99	0.99	5817
Overall Accuracy			0.99	168224

Table B8. Confusion Matrix of High Heart Rate Events Random Forest Model tested on INCART Database

		Prediction	
		0	1
Actual	0	162373	34
	1	16	5801

Table B9. Classification Report of High Heart Rate Events Random Forest Model tested on MIT-BIH Arrhythmia Database

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	46493
Risk	1.00	1.00	1.00	40639
Overall Accuracy			1.00	87132

Table B10. Confusion Matrix of High Heart Rate Events Random Forest Model tested on INCART Database

		Prediction	
		0	1
Actual	0	46493	0
	1	0	40639

Table B11. Classification Report of Resting Heart Rate Logistic Regression Model tested on INCART Database

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	958
Risk	1.00	1.00	1.00	1187
Overall Accuracy			0.99	2145

Table B12. Confusion Matrix of Resting Heart Rate Logistic Regression Model tested on INCART Database

		Prediction	
		0	1
Actual	0	957	1
	1	0	1187

Table B13. Classification Report of Resting Heart Rate Logistic Regression Model tested on MIMIC-III Database

Class	Precision	Recall	F1-Score	Support
No Risk	1.00	1.00	1.00	96
Risk	1.00	1.00	1.00	218
Overall Accuracy			1.00	314

Table B14. Confusion Matrix of Resting Heart Rate Logistic Regression Model tested on MIMIC-III Database

		Prediction	
		0	1
Actual	0	96	0
	1	0	218

Table B15. Classification Report of Heart Rate Variability XGBoost Model tested on MIT-BIH Arrhythmia Database

Class	Precision	Recall	F1-Score	Support
No Risk	0.89	0.69	0.78	58217
Risk	0.34	0.66	0.45	15970
Overall Accuracy			0.72	74187

Table B16. Confusion Matrix of Heart Rate Variability XGBoost Model tested on MIT-BIH Arrhythmia Database

		Prediction	
		0	1
Actual	0	40190	18027
	1	5460	10510

Table B17. Classification Report of Heart Rate Variability LightGBM Model tested on MIT-BIH Arrhythmia Database

Class	Precision	Recall	F1-Score	Support
No Risk	0.73	0.94	0.82	75277
Risk	0.35	0.75	0.48	14581
Overall Accuracy			0.74	89858

Table B18. Confusion Matrix of Heart Rate Variability LightGBM Model tested on MIT-BIH Arrhythmia Database

		Prediction	
		0	1
Actual	0	45934	12283
	1	5450	10520

Table B19. Classification Report of Heart Rate Variability CatBoost Model tested on MIT-BIH Arrhythmia Database

Class	Precision	Recall	F1-Score	Support
No Risk	0.95	0.68	0.79	65748
Risk	0.14	0.58	0.23	5898
Overall Accuracy			0.72	71646

Table B20. Confusion Matrix of Heart Rate Variability CatBoost Model tested on MIT-BIH Arrhythmia Database

		Prediction	
		0	1
Actual	0	52831	22094
	1	3207	11307

Table B21. Classification Report of Heart Rate Variability Logistic Regression Ensemble Model tested on MIT-BIH Arrhythmia Database

Class	Precision	Recall	F1-Score	Support
No Risk	0.90	0.66	0.76	90230
Risk	0.33	0.70	0.45	21889
Overall Accuracy			0.67	71646

Table B22. Confusion Matrix of Heart Rate Variability Logistic Regression Ensemble Model tested on MIT-BIH Arrhythmia Database

		Prediction	
		0	1
Actual	0	59494	30736
	1	6568	15321

Table B23. Classification Report of Heart Rate Variability Logistic Regression Ensemble Model tested on INCART Database

Class	Precision	Recall	F1-Score	Support
No Risk	0.97	0.69	0.81	36781
Risk	0.13	0.70	0.22	2454
Overall Accuracy			0.69	39235

Table B24. Confusion Matrix of Heart Rate Variability Logistic Regression Ensemble Model tested on INCART Database

		Prediction	
		0	1
Actual	0	25423	11358
	1	724	1730

Table B25. Classification Report of Electrocardiogram BiLSTM Model tested on MIT-BIH Arrhythmia combined with Holter Database

Class	Precision	Recall	F1-Score	Support
Low Risk	0.99	0.99	0.99	24506
Medium Risk	0.73	1.00	0.84	476
High Risk	0.99	0.99	0.99	23924
Overall Accuracy			0.98	48906

Table B26. Confusion Matrix of Electrocardiogram BiLSTM Model tested on MIT-BIH Arrhythmia combined with Holter Database

		Prediction		
		Low	Medium	High
Actual	Low	24177	75	254
	Medium	0	475	1
	High	170	104	23650

Table B27. Classification Report of Electrocardiogram BiLSTM Model tested on INCART Database

Class	Precision	Recall	F1-Score	Support
Low Risk	0.96	0.60	0.74	153596
Medium Risk	0.03	0.69	0.06	1959
High Risk	0.45	0.78	0.57	20225
Overall Accuracy			0.62	175780

Table B28. Confusion Matrix of Electrocardiogram BiLSTM Model tested on INCART Database

		Prediction		
		Low	Medium	High
Actual	Low	91664	43100	18832
	Medium	545	1355	59
	High	3473	1001	15751

Table B29. Classification Report of Electrocardiogram BiLSTM Model tested on OHCA Database

Class	Precision	Recall	F1-Score	Support
Low Risk	0.00	0.00	0.00	0
Medium Risk	0.00	0.00	0.00	0
High Risk	1.00	0.97	0.98	3574
Overall Accuracy			0.96	3574

Table B30. Confusion Matrix of Electrocardiogram BiLSTM Model tested on OHCA Database

		Prediction		
		Low	Medium	High
Actual	Low	0	0	0
	Medium	0	0	0
	High	80	39	3455

Table B31. Classification Report of Final Ensemble Fine-Tuned BiLSTM Model tested on MIT-BIH Arrhythmia combined with Holter Database

Class	Precision	Recall	F1-Score	Support
Low Risk	0.99	0.99	0.99	24506
Medium Risk	0.89	1.00	0.94	476
High Risk	0.99	0.99	0.99	23924
Overall Accuracy			0.99	48906

Table B32. Confusion Matrix of Final Ensemble Fine-Tuned BiLSTM Model tested on MIT-BIH Arrhythmia combined with Holter Database

		Prediction		
		Low	Medium	High
Actual	Low	24200	60	246
	Medium	0	475	1
	High	60	40	23824

Table 33. Classification Report of Final Ensemble Fine-Tuned BiLSTM Model tested on INCART Database

Class	Precision	Recall	F1-Score	Support
Low Risk	0.96	0.69	0.80	153596
Medium Risk	0.05	0.71	0.09	476
High Risk	0.47	0.83	0.60	20225
Overall Accuracy			0.67	175780

Table B34. Confusion Matrix of Final Ensemble Fine-Tuned BiLSTM Model tested on INCART Database

		Prediction		
		Low	Medium	High
Actual	Low	105679	29417	18500
	Medium	500	1400	59
	High	2500	1000	16725

Table B35. Classification Report of Final Ensemble Fine-Tuned BiLSTM Model tested on OHCA Database

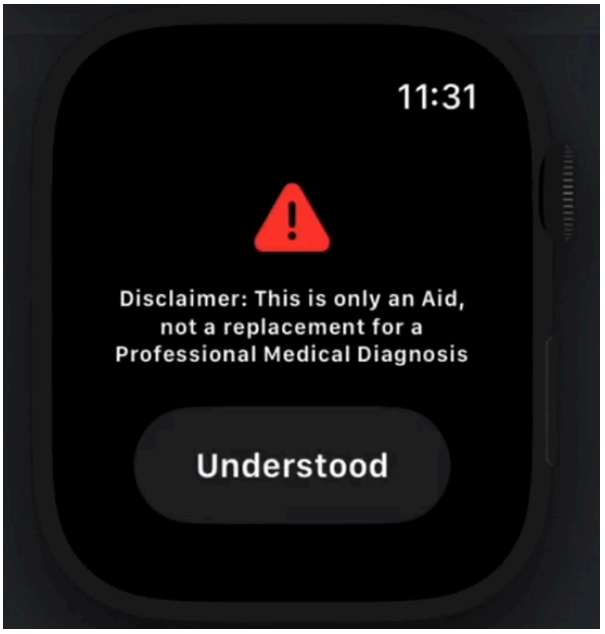
Class	Precision	Recall	F1-Score	Support
Low Risk	0.00	0.00	0.00	0
Medium Risk	0.00	0.00	0.00	0
High Risk	0.99	0.98	0.98	3574
Overall Accuracy			0.98	3574

Table B36. Confusion Matrix of Final Ensemble Fine-Tuned BiLSTM Model tested on OHCA Database

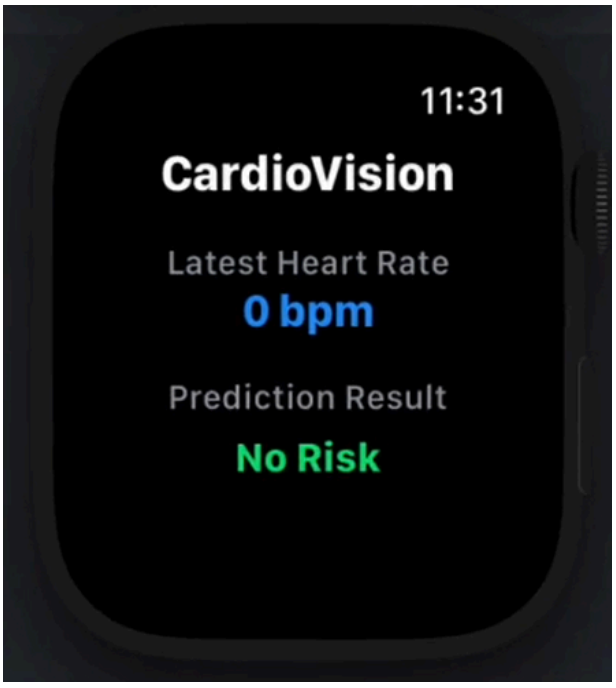
		Prediction		
		Low	Medium	High
Actual	Low	0	0	0
	Medium	0	0	0
	High	30	20	3524

Appendix C. WatchOS App Screens

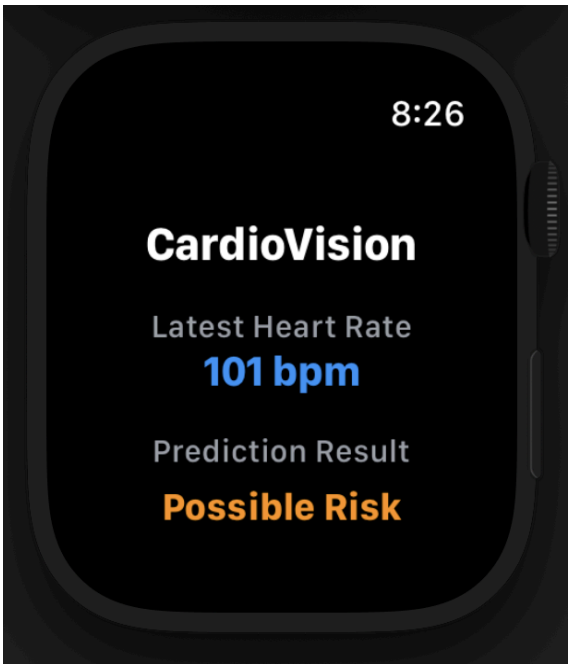
Screen C1. Disclaimer Screen



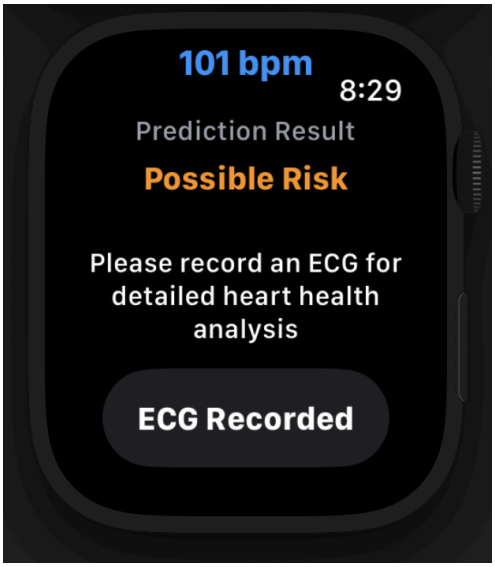
Screen C2. Initial No Risk Monitoring Screen

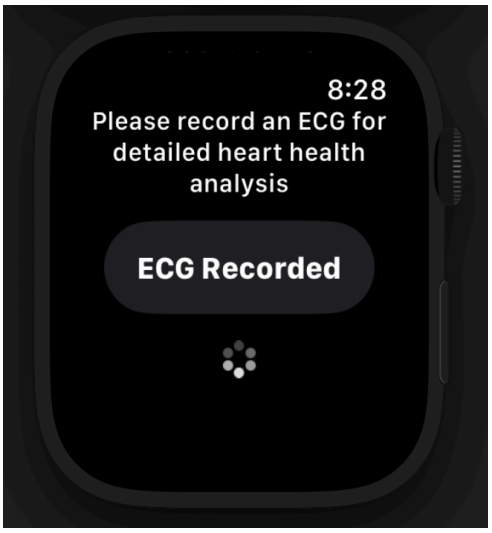


Screen C3. Initial Risk Detected & Prompt ECG Screen



Screen C4. ECG Analysis Screen





Screen C5. Final Low Risk Detected Screen

Screen C6. Final Medium Risk Detected Screen

Screen C7. Final High Risk Detected Screen