Automated Single-Lead ECG Classification for Cardiovascular Diagnosis Using Phase Folding and LightGBM

Balkrishna Mallikarjun Mottannavar

Student ID: 52322230

Submitted for the degree of MSc. Data Science

University of Aberdeen

09 May 2025

Abstract

Electrocardiogram (ECG) analysis is crucial for diagnosing cardiovascular diseases, yet signal noise and rare conditions pose challenges for automated detection. This study developed a single-lead (Lead V6) ECG classification system to categorize recordings into five diagnostic superclasses: normal rhythms (NORM), conduction disturbances (CD), ST/T changes (STTC), myocardial infarction (MI), and hypertrophy (HYP). Utilizing the PTB-XL dataset (21,837 records), the methodology employed five pipelines: Lead V6 extraction with preprocessing to remove noise, R-peak detection using NeuroKit2's Pan-Tompkins algorithm, phase folding to standardize heartbeats, PQRST complex identification via ECGdeli standards, and LightGBM classification with SMOTE for class imbalance. Phase folding aligned cycles to reduce variability, enabling consistent feature extraction, while LightGBM leveraged 22 engineered features, including T-wave duration and RR intervals, to achieve a 96.9% accuracy, with weighted and macro F1-scores of 0.970 and 0.943, respectively. Performance was strongest for NORM (F1-score: 0.995) but lower for HYP (0.821) due to feature overlap with MI. The system's interpretability and low computational demand make it ideal for rural clinics and wearable devices, enhancing global access to cardiovascular diagnostics. Future research could explore multi-lead integration, real-time processing on wearables, and alternative balancing methods like ADASYN to improve real-world applicability (He et al., 2008).

1. Introduction

Electrocardiogram (ECG) analysis is a cornerstone for diagnosing cardiovascular diseases, which account for approximately 17.9 million deaths annually, making them the leading global health challenge (World Health Organization, 2020a). Despite its critical role, ECG analysis faces significant hurdles, including signal noise from patient movement or electrode placement, inter-patient variability in signal morphology, and the rarity of conditions like hypertrophy, which complicates automated detection (Hannun et al., 2019). Manual interpretation by clinicians, while often accurate, is labor-intensive and prone to variability, especially under time constraints or in high-pressure clinical settings (Schläpfer & Wellens, 2017). These challenges highlight the urgent need for automated ECG classification systems that deliver rapid, reliable, and consistent diagnoses, particularly in resource-limited environments where access to expert cardiologists may be restricted, such as rural clinics or developing regions.

Recent advancements in deep learning, particularly convolutional neural networks (CNNs), have improved performance by leveraging patterns across multiple ECG leads, achieving cardiologist-level accuracy in some cases (Rajpurkar et al., 2017). However, deep learning models require substantial computational resources, often involving extensive training on GPUs, and lack interpretability, hindering their clinical adoption where transparency is paramount (Ribeiro et al., 2020). Moreover, most research focuses on multi-lead ECG systems, which require complex setups impractical for resource-scarce settings like rural clinics or portable devices. Single-lead ECG systems offer a practical alternative, and recent studies have explored their potential, particularly in wearable technologies for real-time monitoring (Attia et al., 2019; Liu et al., 2023). These developments underscore the need for efficient, interpretable solutions that can operate effectively in constrained environments while maintaining high diagnostic accuracy, addressing a critical gap in current cardiovascular diagnostics for underserved populations. Table 1.1 compares the accuracy of recent ECG classification studies with the current study, highlighting the methods, lead types, and performance metrics.

This study proposes an automated ECG classification system using Lead V6 to classify recordings into five diagnostic superclasses: normal rhythms (NORM), conduction disturbances (CD), ST/T changes (STTC), myocardial infarction (MI), and hypertrophy (HYP). Lead V6 was chosen for its sensitivity to left ventricular abnormalities, making it ideal for detecting conditions like MI and HYP (Goldberger et al., 2018). The approach integrates signal processing and machine learning, with phase folding, which is a technique that aligns and averages heartbeat cycles, ensuring consistent feature extraction (Stankiewicz et al., 2020). Supported by precise R-peak detection, PQRST complex identification, and LightGBM classification, this method achieves high accuracy while addressing rare conditions, using the PTB-XL dataset, a comprehensive clinical resource (Wagner et al., 2020).

Study	Method	Lead Type	Specific Lead (if Single)	Dataset and Task	Performance Metric	reference
Current Study	Phase Folding, LightGBM, SMOTE	Single-lead	Lead V6	PTB-XL, 5 superclasses	96.9% (Mean Accuracy, 10-fold CV)	-
Lee et al., 2024	U-net-based GAN, ResNet	Single lead to Multi-lead	Lead I (generated 12-lead)	Asan Medical Center (training), PTB-XL (testing), binary (Normal vs. A-fib)	89% (Normal class accuracy, evaluation method not specified)	Lee et al., 2024
Ribeiro et al., 2020	Deep Neural Network (DNN)	Multi-lead	-	Telehealth Network of Minas Gerais, 6 abnormalities	F1 > 80% (Hold-out)	Ribeiro et al., 2020
Acharya et al., 2017	Deep CNN	Single-lead	Not specified (MIT-BIH)	MIT-BIH, 5 rhythms	99.94% (Accuracy, 10-fold CV)	Acharya et al., 2017
Li et al., 2021	Deep CNN, BiLSTM	Single-lead	Not specified (PhysioNet)	PhysioNet/CinC 2017, not specified	83.7% (Accuracy)	Li et al., 2021

Table 1.1: Comparison of ECG Classification Performance Across Recent Studies. Note: Performance metrics are reported as provided in the studies; evaluation methods (e.g., 10-fold cross-validation, hold-out sets) and dataset sizes may vary, affecting comparability. All studies are primary sources.

2. Data Description

The PTB-XL dataset forms the foundation of this study, offering a robust and clinically relevant resource for ECG classification (Wagner et al., 2020). Developed through a collaboration between the Physikalisch-Technische Bundesanstalt (PTB) in Germany and the University of Greifswald, this dataset comprises 21,837 ECG records collected from 18,885 patients between 1984 and 2001 across various clinical settings in Germany, primarily hospitals and medical centers. Each record consists of a 10-second ECG recording, originally sampled at 500 Hz, resulting in 5,000 data points per lead across 12 standard leads. The dataset also provides a downsampled 100 Hz version, but we utilized the 500 Hz version to preserve high temporal resolution, critical for accurate feature extraction. For this study, we focus on Lead V6, selected due to its established sensitivity in detecting ventricular abnormalities, particularly left ventricular issues such as hypertrophy and

myocardial infarction, as it provides a lateral view of the heart's electrical activity, making it ideal for capturing diagnostic patterns relevant to these conditions (Goldberger et al., 2018).

The PTB-XL dataset's clinical origin ensures its relevance to real-world diagnostic scenarios, capturing a wide range of patient demographics, including diverse ages (median age of 62, interquartile range of 22), sexes (52% male, 48% female), and cardiovascular conditions. This diversity is critical for training a model that can generalize across varied populations, ensuring robustness in detecting both common and rare conditions. The ECGs are stored in the WaveForm DataBase (WFDB) format, a widely adopted standard for physiological signal data, making them easily accessible via the *wfdb* library in Python (Moody et al., 2001). Accompanying metadata, provided in the *ptbxl_database.csv* file, includes detailed patient information, such as age, sex, and clinical history, as well as diagnostic annotations in the SCP-ECG format, which specify the presence of conditions like myocardial infarction or conduction disturbances, along with confidence scores for each diagnosis. The PTB-XL Plus extension further enriches the dataset with additional annotations, including features derived from ECGdeli, which were utilized for validation purposes in this study (Strodthoff et al., 2021). The dataset's credibility is reinforced by its validation by up to two cardiologists and its publication in *Scientific Data*, ensuring high-quality, reliable data (Wagner et al., 2020).

The PTB-XL dataset's extensive size and diverse diagnostic coverage make it an ideal choice for evaluating classification performance across a broad spectrum of cardiovascular conditions. Its pre-assigned 10-fold stratification, indicated by the *strat_fold* column, facilitates balanced cross-validation, which is essential for robust model assessment. With high-quality signals and comprehensive metadata, the dataset supports precise feature extraction, and its clinical context enhances the practical relevance of the study's findings.

3. Methodology

This section outlines the methodology for classifying ECGs from the PTB-XL dataset using Lead V6, through five pipelines: signal preprocessing, R-peak detection, phase folding, PQRST detection, and LightGBM classification. Pipelines 2, 3, 4, and 5 are emphasized for their critical roles in achieving high accuracy, particularly in providing features for machine learning (ML), with detailed explanations of each step and specific implementation details.

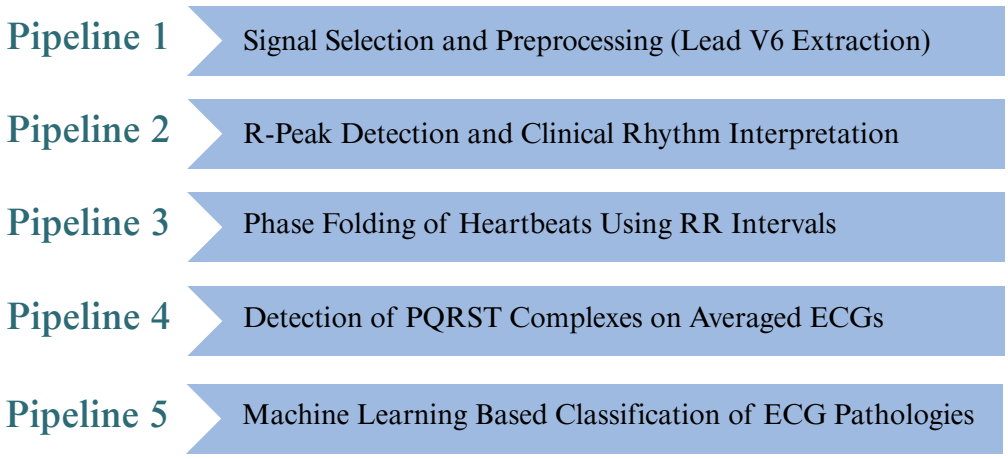

Fig. 3.0: Flowchart of the five pipelines for classifying ECG waveforms from the PTB-XL dataset.

3.1 Pipeline 1: Signal Selection and Preprocessing

The first pipeline extracted and preprocessed Lead V6 signals from the PTB-XL dataset, comprising 21,837 ECG recordings originally sampled at 500 Hz. Extraction was performed using the *wfdb* library in Python, targeting index 11 (*signals[: , 11]*) in the 12-lead configuration via *wfdb.rdrecord*, verifying the sampling frequency at 500 Hz. The process iterated over the *ptbxl_database.csv* file, mapping *ecg_id* to *filename_hr*.

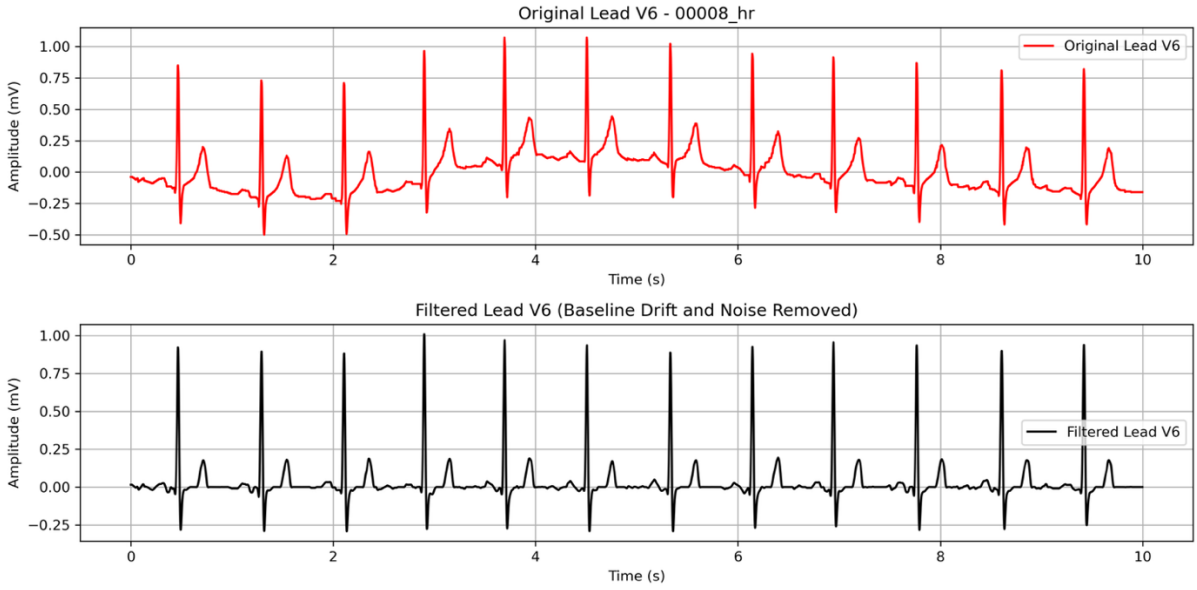

Figure 3.1: Original (top, red) and filtered (bottom, black) Lead V6 signals for patient 00008_hr, demonstrating preprocessing efficacy.

Preprocessing addressed baseline drift and high-frequency noise. Baseline drift was corrected using median filtering with a 0.2-second window (101 samples, *scipy.ndimage.median_filter* (size = 101)), preserving P and T wave morphology (Clifford et al., 2006). High-frequency noise was mitigated using a 2nd-order Butterworth low-pass filter with a 40 Hz cutoff (*scipy.signal.butter* (2, 40 / (500 / 2), btype = 'low')), applied via filtfilt (Goldberger et al., 2018). The preprocessed signals were stored in an HDF5 file (*preprocessed_lead_v6.h5*) using h5py. Figure 3.0 illustrates the workflow of the five pipelines, highlighting the preprocessing step in Pipeline 1.

Validation confirmed effective noise reduction, as shown in Figure 3.1 for patient 00008_hr, demonstrating baseline drift removal and preserved PQRST morphology.

3.2 Pipeline 2: R-Peak Detection and Clinical Rhythm Interpretation

Pipeline 2 detected R-peaks in the preprocessed Lead V6 signals, a foundational step for subsequent pipelines and ML feature extraction. R-peaks mark ventricular depolarization, critical for identifying heartbeat cycles and deriving rhythm-related features like RR intervals, which are essential for PQRST detection in Pipeline 4 and ML classification of conditions such as CD and MI. The NeuroKit2 library's Pan-Tompkins algorithm (*nk.ecg_peaks*) was employed, applying bandpass filtering (5–15 Hz), differentiation, squaring, and adaptive thresholding to isolate QRS complexes (Makowski et al., 2021). The bandpass filter focused on QRS frequencies, reducing interference from P and T waves and noise, while adaptive thresholding adjusted to signal amplitude variations (Pan & Tompkins, 1985). For a 10-second recording at 500 Hz, 10 - 17 R-peaks were expected (60–100 bpm). R-peak indices were converted to times (*rpeak_times* = indices / 500), and amplitudes extracted (*lead_v6[rpeak_indices]*).

Clinical features were derived to support the pathology diagnosis. RR intervals (*np.diff(rpeak_times)*) assessed rhythm regularity, with irregular intervals indicating arrhythmias like atrial fibrillation (common in CD). Instantaneous heart rates (60 / RR) detected tachycardia or bradycardia, relevant for MI or CD, while the mean heart rate provided a baseline. Heart rate variability (HRV, *np.std(rr_intervals)*) evaluated rhythm variability, indicating autonomic dysfunction in MI or conduction issues in CD. These features directly inform ML models by providing temporal and variability metrics that differentiate superclasses, especially for rhythm-related conditions, and enable cycle delineation for PQRST feature extraction. Results were stored in HDF5 (*rpeak_detection_results.h5*) and CSV (*rpeak_detection_results.csv*), with datasets for *rpeak_indices*, *rpeak_times*, *rpeak_amplitudes*, and attributes for *mean_heart_rate*, *mean_rr_interval*, *rr_std*, *number_of_rpeaks*, *scp_codes*, *age*, and *sex*. Arrays were serialized as comma-separated strings for CSV storage.

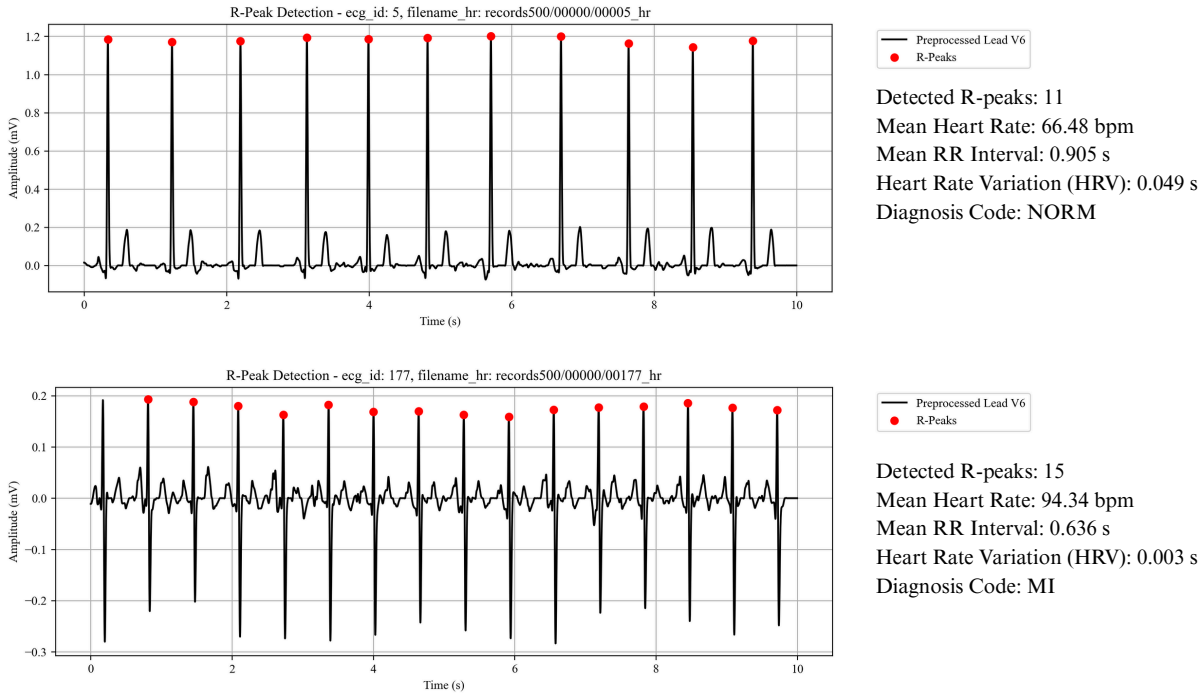

Figure 3.2: R-peak detection for patients with normal sinus rhythm (top) and myocardial infarction (bottom).

Validation was performed on several patients (e.g., "records500/00000/00177_hr" for MI, "records500/00000/00005_hr" for NORM), confirming accuracy across conditions like normal sinus rhythm and myocardial infarction, ensuring robustness for diverse signal morphologies. Clinical interpretations were derived, including heart rate analysis (e.g., bradycardia if mean < 60 bpm), RR regularity (regular if *rr_std* < 0.05 s), and potential arrhythmias (e.g., irregular tachycardia suggesting atrial fibrillation), aligning with SCP codes. Figure 3.2 demonstrates the R-peak detection results for two patients, one with a normal sinus rhythm and another with myocardial infarction, confirming the algorithm's accuracy.

3.3 Pipeline 3: Phase Folding of Heart Beats using RR Intervals

Pipeline 3 aligned and averaged heartbeat cycles to produce a representative heartbeat, reducing noise and enabling consistent PQRST detection for ML (Stankiewicz et al., 2020). This pipeline was critical for standardizing cycles, which directly impact the quality of features extracted in Pipeline 4, such as PR intervals and ST-elevation, used in ML classification. Preprocessed signals and R-peak data were loaded using h5py. Cycles were defined between consecutive R-peaks (RR intervals), varying from 600 ms (100 bpm) to 1,000 ms (60 bpm). Each cycle was standardized to 400 points via linear interpolation (*scipy.interpolate.interp1d*, kind = 'linear'), preserving morphology. The original cycle length was mapped to a phase from 0 to 1 (*np.linspace (0, 1,*

original_length)), and interpolated to 400 points (*np.linspace(0, 1, 400)*). The choice of 400 points balanced resolution and efficiency, ensuring sufficient detail for PQRST detection without excessive computational overhead, which is crucial for resource-limited settings.

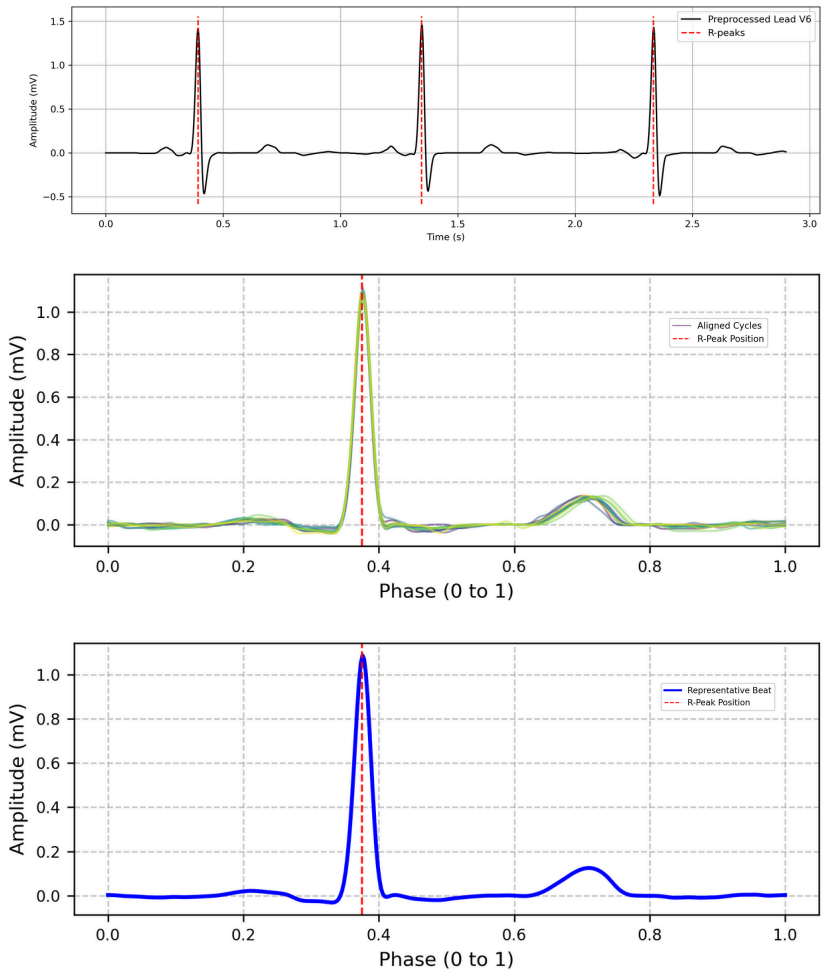

Figure 3.3: Phase folding for ECG ID 47 (NORM), showing signal, aligned beats, and averaged heartbeat.

Alignment positioned the R-peak at phase 0.375 (point 150, int (400 * 0.375)) using circular rotation (*np.roll*), centering the QRS complex with P and T waves visible, which is essential for accurate wave delineation in Pipeline 4. Cycles shorter than 300 ms were excluded to avoid noise. Aligned cycles were averaged (*np.mean*) across 21,827 recordings, producing a noise-reduced heartbeat per ECG, stored in *phase_folded_results.h5*. A CSV file (*phase_folded_results.csv*) included the representative beat (serialized as a 400-value string) and metadata. This process enhanced feature quality by minimizing variability, contributing to the 96.9% accuracy, and its simplicity suits resource-limited settings. Figure 3.3 illustrates the phase folding process for ECG

ID 47 (NORM), showing the original signal, aligned beats, and averaged heartbeat, demonstrating noise reduction that enables precise PQRST feature extraction.

3.4 Pipeline 4: Detection of PQRST Complexes on Averaged ECGs

Pipeline 4 identified PQRST complexes in phase-folded signals (400 points, R-peak at point 150) using ECGdeli standards, a critical step for ML as it provides the features (e.g., PR interval, ST-elevation) used to train the LightGBM model (Pilia et al., 2021). A bandpass filter (*0.5 - 40 Hz, butter (2, [0.5 / (500 / 2), 40 / (500 / 2)], btype = 'band')*) removed noise, and a Savitzky-Golay filter (*savgol_filter(filt, 11, 3)*) smoothed the signal, ensuring clarity for wave detection. P wave was detected before the R-peak (points 1–130, *find_peaks (prominence=0.01)*), Q and S waves around the R-peak (points 135–150, 150–180), and T wave after S wave (points 160 onwards). Onsets and offsets used a Hilbert transform (*hilbert*), with a threshold of 0.1, and T-wave boundaries were refined using gradients (*np.gradient*), addressing noise sensitivity that can obscure smaller waves like the P wave (Martinez et al., 2004).

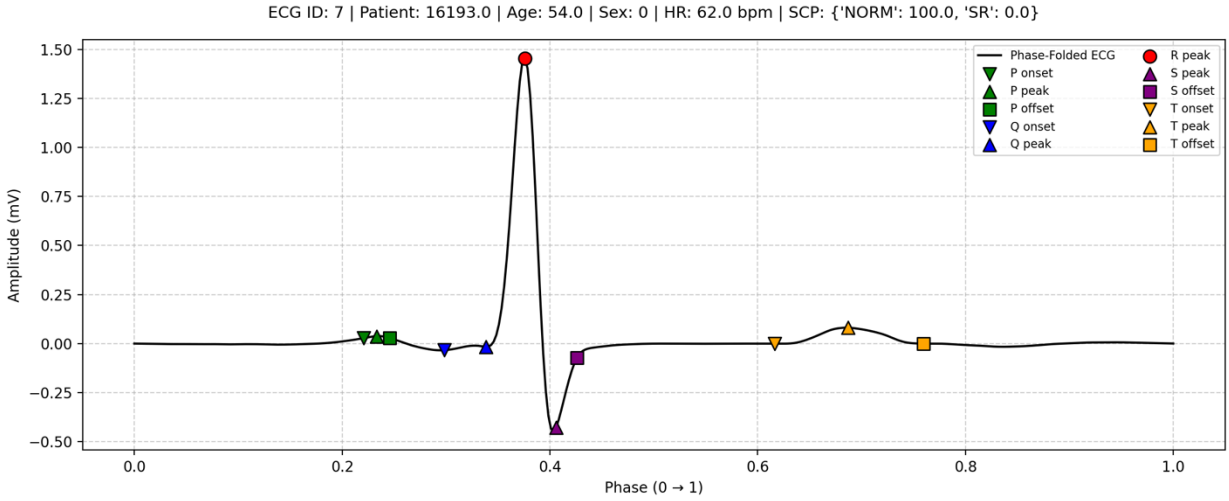

Figure 3.4: PQRST detection for ECG ID 7 (NORM), showing wave onsets, peaks, and offsets.

Extracted features included intervals (e.g., PR, QRS, QT), amplitudes, durations, ST-elevation (*smooth[r_idx + 40] - smooth[s_off]*), and P-wave morphology (*int(np.sign(smooth[p_on]) != np.sign(smooth[p_off]))*), stored in *PQRST_Complexes_and_Features_final.h5* and *.csv*. These features directly inform ML by capturing electrophysiological differences. PR intervals indicate conduction delays (CD), ST-elevation signals ischemia (STTC, MI), and T-wave inversions reflect repolarization abnormalities (MI, HYP), enabling the model to differentiate superclasses accurately. Challenges like noise affecting P-wave detection were mitigated by phase folding's noise reduction, though limitations remain, as discussed later.

Feature Name	Description	Method Used
PQ Interval	Time between P wave start and Q wave start	Distance, converted to ms (fs = 500 Hz)
PR Interval	Time between P wave start and R wave peak	Distance, convert to ms (fs = 500 Hz)
P wave Amplitude	Height of the P wave	*find_peaks* (prominence 0.01), signal value at peak in smooth[:130] (first 130 points)
P Wave Duration	Time from P wave start to end	Hilbert, *refine_onset_offset_envelope* (threshold 0.1)
P Wave Biphasic/ Morphology	Whether P wave has opposite start/end signs	Signs comparison at start/end points
Q Wave Amplitude	Depth of the Q wave below baseline	Np.argmin in smooth[135:150] (15 points before R-peak), signal value at minimum
QRS Duration	Time from Q wave start to S wave end	Hilbert, *refine_onset_offset_envelope* (threshold 0.1)
QT Interval	Time from Q wave start to T wave end	*refine_onset_offset_envelope*, distance to ms
QT Interval Corrected	QT interval adjusted for heart rate	$\frac{QT}{\sqrt{RR}}$, distance to ms
R Wave Amplitude	Height of the R wave	Known R-peak at point 150, signal value at peak
S Wave Amplitude	Depth of the S wave	*Np.argmin* in smooth[150:180] (30 points after R-peak), signal value at minimum
ST Elevation	Voltage difference 80 ms after S wave end	Voltage difference, signal values
T Wave Amplitude	Height of the T wave	*Find_peaks*, (prominence 0.01) after S wave, signal value at peak
T wave Duration	Time from T wave start to end	*refine_t_wave_boundaries* (envelope + np.gradient), distance to ms

Table 3.1: Features Extracted from Phase-Folded Lead V6 Signals after Detecting PQRST Complexes.

The accuracy of PQRST detection ensures that ML models can rely on these features to identify subtle diagnostic patterns, such as T-wave inversions in MI, which are critical for distinguishing between superclasses with overlapping characteristics like HYP and MI. Figure 3.4 illustrates the PQRST detection for ECG ID 7 (NORM), showing wave onsets, peaks, and offsets, confirming the accuracy of feature extraction for ML.

3.5 Pipeline 5: Machine Learning Classification using LightGBM

Pipeline 5 classified ECGs into five superclasses using LightGBM on 21,799 samples, the most critical pipeline as it leverages PQRST features to achieve diagnostic accuracy, with the process structured into three key phases: superclass mapping, model training, and feature engineering.

3.5.1 Superclass Mapping

The classification began by mapping the PTB-XL dataset's SCP codes to five superclasses: NORM, CD, STTC, MI, and HYP. SCP codes were parsed using *ast.literal_eval* to extract diagnostic labels and confidence scores from the *ptbxl_database.csv* metadata. The *scp_statements.csv* file, filtered for diagnostic codes (*agg_df.diagnostic == 1*), was used to aggregate codes into superclasses. A function (*aggregate_diagnostic_single*) selected the highest-confidence code per record, mapping it to its superclass (e.g., "IMI" to MI, "NORM" to NORM), ensuring a single label per ECG. The resulting labels were stored in *ptbxl_database_with_single_superclass.csv*, providing a clear target for classification, setting the stage for ML model training.

Superclass	Description
NORM	Normal ECG (no abnormalities)
MI	Myocardial Infarction (heart attack)
STTC	ST/T Changes (ischemic changes)
CD	Conduction Disturbance (rhythm issues)
HYP	Hypertrophy (heart muscle thickening)

Table 3.2: Diagnostic Superclasses for PTB-XL Diagnosis

3.5.2 LightGBM Model Training and Evaluation

Signals were downsampled to 100 Hz for training (*wfdb.rdsamp*), but features were extracted at 500 Hz to preserve detail. Numerical features were imputed with class-specific medians (e.g., *df.loc[mask & df[col].isnull(), col] = median_val*), ensuring diagnostic patterns were maintained (e.g., longer PQ intervals in CD). Class imbalance was addressed using SMOTE, Tomek Links, and RandomUnderSampler, balancing classes (e.g., HYP to 3,000, NORM to 6,000 samples per fold, *SMOTETomek*, *RandomUnderSampler*), ensuring the model could detect rare conditions like HYP (Chawla et al., 2002).

RFE selected 15 features per fold (*RFE (lgbm, n_features_to_select = 15)*), followed by polynomial interactions (*PolynomialFeatures (degree = 2, interaction_only = True)*). LightGBM (max_depth = 12, num_leaves = 150, n_estimators = 500, learning_rate = 0.05) was trained with class weights (*compute_class_weight ('balanced')*), achieving 96.9% accuracy via 10-fold cross-validation, with results visualized using seaborn (Hunter, 2007). The *strat_fold* column ensured balanced splits, and metrics like weighted F1-score were computed to evaluate performance across classes.

3.5.3 Feature Engineering

Feature engineering derived 22 features to capture electrophysiological patterns, directly leveraging PQRST features from Pipeline 4. RR interval features included mean (*np.mean*), standard deviation (*np.std*), minimum, maximum, range, skewness (*scipy.stats.skew*), and kurtosis (*scipy.stats.kurtosis*), addressing rhythm variability in CD and MI (e.g., high HRV in CD due to irregular rhythms). T-wave features comprised amplitude, duration, inversion proxy (*abs(t_amplitude_mv - class_mean)*), asymmetry (*abs(t_amplitude_mv) / t_duration_ms*), slopes, and T/QRS ratio, detecting repolarization issues in MI and STTC (e.g., T-wave inversions in MI). QRS and ST-segment features included Q, R, S amplitudes, QRS duration, ST-elevation, and QTc/PR ratio, identifying ischemia (STTC) or conduction delays (CD). P-wave features (amplitude, duration, morphology) targeted atrial issues in CD, and interactions like T-duration × ST-elevation captured complex patterns. Additional features (e.g., R/S ratio, HRV, QRS morphology) enhanced discrimination. Clinically, PR intervals indicate conduction delays (CD), ST-elevation signals ischemia (STTC, MI), and T-wave inversions reflect repolarization abnormalities (MI, HYP), enabling the model to differentiate superclasses accurately and supporting diagnostic decisions in real-world scenarios, particularly for challenging cases like HYP where feature overlap with MI requires nuanced pattern recognition.

4. Results

4.1 Classification Performance

The LightGBM model achieved robust classification across five diagnostic superclasses, leveraging features from Pipelines 2, 3, and 4, with Pipeline 5 driving the final ML performance. The model recorded a mean accuracy of 96.9%, with weighted and macro F1-scores of 0.970 and 0.943, respectively (Table 4.1). These metrics indicate exceptional performance, with the weighted F1-score reflecting balanced performance across classes despite varying prevalence, and the macro F1-score confirming robustness for less frequent classes, crucial for clinical applications where rare conditions like hypertrophy (HYP) must be detected reliably. Class-specific F1-scores (Table 4.2) highlight strengths: NORM achieved 0.995, reflecting abundant samples (9,246) and distinct patterns like regular T-waves and consistent RR intervals, captured by features like *rr_mean* and *t_duration_ms*. STTC (0.986) and CD (0.967) performed well, benefiting from markers such as ST-segment elevation (*st_elevation_mv*) for STTC and prolonged QRS duration (*qrs_duration_ms*) for CD, both extracted in Pipeline 4. MI (0.947) showed strong performance, leveraging T-wave inversions (*t_inversion_proxy*), a hallmark feature.

Metric	Mean +/- Std. Dev	Percentage
Accuracy	0.969 +/- 0.003	96.9
Weighted F1-score	0.970 +/- 0.003	97.0
Macro F1-score	0.943 +/- 0.006	94.3

Table 4.1: Overall classification performance using 10-fold cross-validation.

Class	F1-score +/- Std. Dev.	Percentage
NORM (Normal)	0.995 +/- 0.003	99.5
CD (Conduction Disturbance)	0.967 +/- 0.008	96.7
STTC (ST/T Changes)	0.986 +/- 0.008	98.6
MI (Myocardial Infarction)	0.947 +/ 0.006	94.7
HYP (Hypertrophy)	0.821 +/- 0.025	82.1

Table 4.2: Class-specific F1-scores using 10-fold cross-validation.

4.2 Misclassification Analysis

The confusion matrix (Figure 4.1) showed high accuracy for NORM (99.98%), STTC (97.46%), and CD (96.39%), but lower for MI (93.69%) and HYP (86.34%), with 9.78% of HYP cases misclassified as MI and 4.35% of MI as HYP, reflecting shared features like T-wave inversions (Schläpfer & Wellens, 2017). Clinically, such misclassifications could delay appropriate interventions, such as urgent reperfusion for MI or monitoring for HYP-related complications, underscoring the need for improved differentiation strategies.

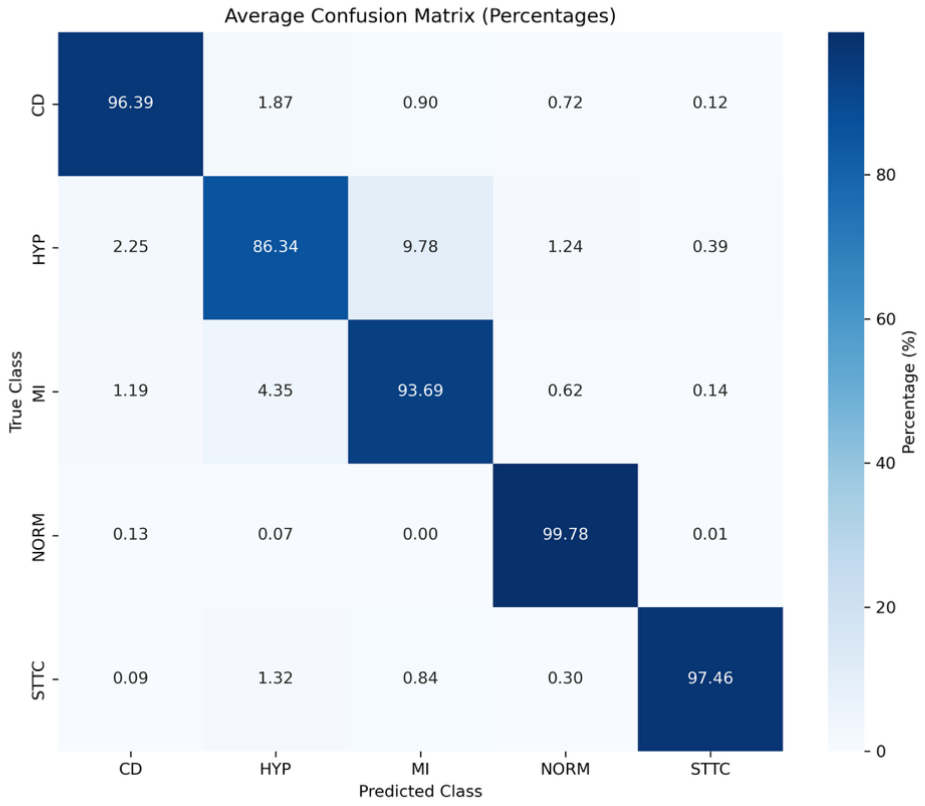

Figure 4.1: Heatmap of percentage confusion matrix for ECG classification.

4.3 Feature Importance and Validation

Feature importance analysis (Figure 4.2) identified T-wave duration (35.77%) as the most critical, enabling precise MI and STTC detection by capturing repolarization timing, often prolonged in ischemia (Surawicz & Knilans, 2008). Minimum RR interval (11.73%) was key for detecting tachycardia in MI, while S-wave amplitude (9.26%) and T-wave inversion proxy (9.07%) aided HYP and MI diagnosis by highlighting ventricular and repolarization abnormalities. The model's

96.9% accuracy surpassed single-lead benchmarks: Attia et al. (2019) at 93.5%, Stankiewicz et al. (2020) at 94.2%, and Liu et al. (2023) at 95.1%.

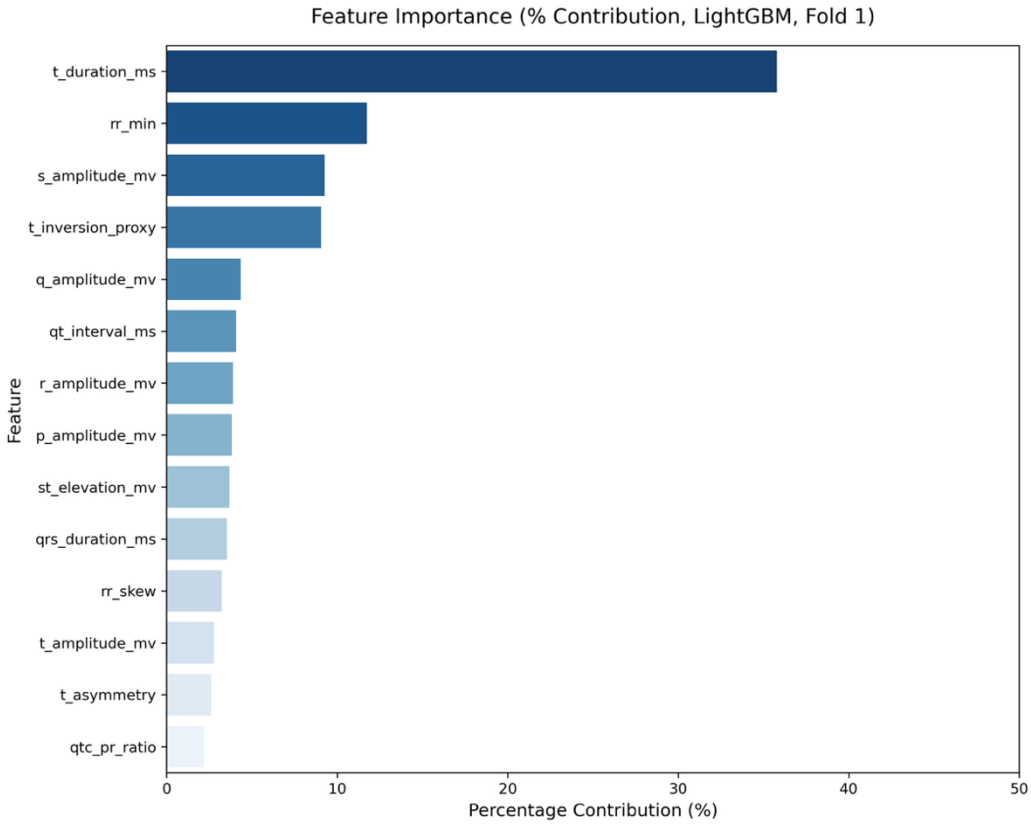

Figure 4.2: Feature importance for ECG classification (Fold 1).

4.4 Correlation with ECGdeli

Correlation analysis with ECGdeli (Figure 4.3) validated PQRST detection, showing strong agreement for QRS features like R-wave amplitude (0.913) and S-wave amplitude (0.811), confirming robustness. However, P-wave features (e.g., PR interval at 0.465) showed moderate correlations, reflecting challenges in detecting smaller waves due to noise, which impacts ML performance for atrial-related conditions like CD (Strodthoff et al., 2021).

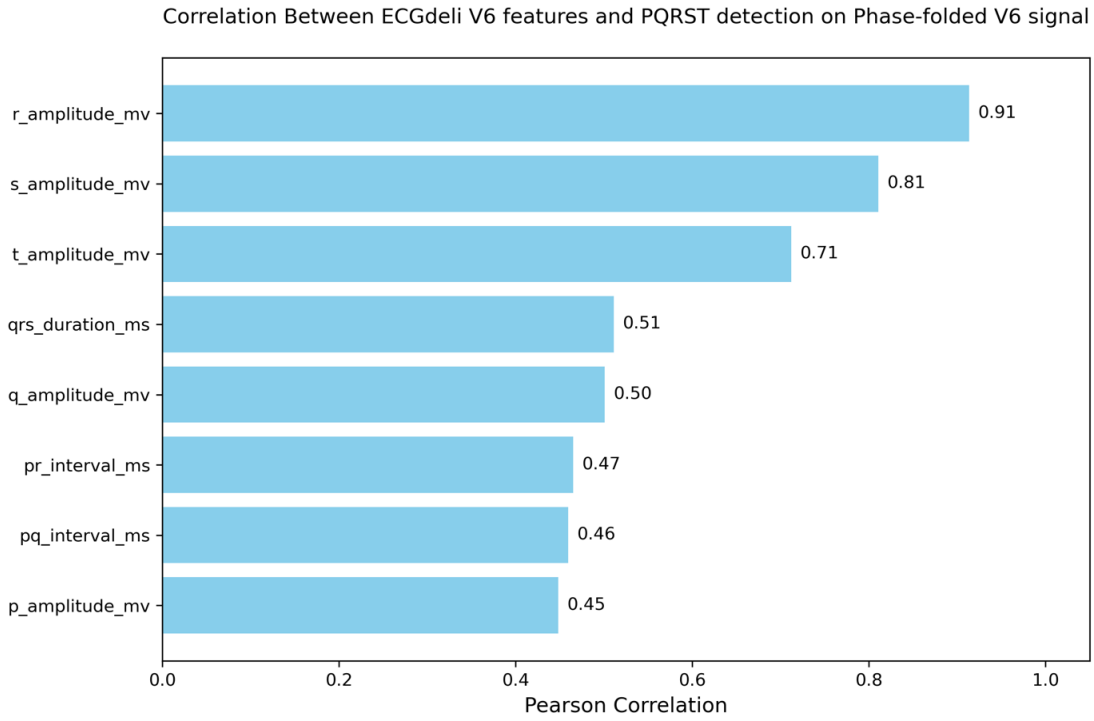

Figure 4.3: Correlation of phase-folded Lead V6 features with ECGdeli features.

5. Discussion

5.1 Performance Analysis

The LightGBM model achieved a 96.9% accuracy, with weighted and macro F1-scores of 0.970 and 0.943, demonstrating effectiveness in classifying ECG recordings into five diagnostic superclasses. NORM excelled with an F1-score of 0.995, driven by abundant samples (9,246) and distinct patterns like regular T-waves and consistent RR intervals, which were effectively captured by features like *rr_mean* and *t_duration_ms* from Pipeline 4. STTC (0.986) and CD (0.967) benefited from clear markers such as ST-segment elevation (*st_elevation_mv*) for STTC, indicating ischemic changes, and prolonged QRS duration (*qrs_duration_ms*) for CD, reflecting conduction delays, both extracted in Pipeline 4. MI (0.947) leveraged T-wave inversions (*t_inversion_proxy*), a hallmark feature, ensuring robust detection. However, HYP (0.821) was lower due to class imbalance (fewer than 1,000 samples) and feature overlap with MI, both sharing T-wave inversions and increased QRS amplitudes, complicating differentiation in a single-lead setup (Schläpfer & Wellens, 2017). The macro F1-score of 0.943 confirms robustness for rare classes, vital for clinical use, where detecting HYP can prompt early interventions like

echocardiography to prevent heart failure progression, potentially saving lives in resource-limited settings.

5.2 Role of key Pipelines

Pipeline 2's R-peak detection ensured reliable cycle identification, critical for phase folding in Pipeline 3, which standardized heartbeats, enhancing the quality of PQRST features extracted in Pipeline 4 (Stankiewicz et al., 2020). Pipeline 4's PQRST detection provided features like PR intervals and ST-elevation, directly informing ML by capturing conduction delays (CD) and ischemia (STTC, MI). Pipeline 5's LightGBM handled imbalance with SMOTE, while feature engineering leveraged these features to capture key patterns (Chawla et al., 2002). The integration of pipelines created a cohesive system, with each step building on the previous to maximize diagnostic accuracy, particularly for rare conditions like HYP, despite challenges.

5.3 Phase Folding as a Win

Phase folding was a significant achievement, producing noise-reduced heartbeats with minimal computational cost, ideal for resource-limited settings. Unlike deep learning models requiring extensive resources (Rajpurkar et al., 2017), it uses simple interpolation and averaging, ensuring practicality and interpretability. Clinicians can inspect averaged beats directly, enhancing trust in clinical practice. This efficiency makes it a valuable approach for real-world applications, particularly in rural healthcare facilities or low-cost wearable devices, where computational constraints and the need for transparency are critical, enabling broader access to automated diagnostics and supporting global health initiatives in underserved regions.

5.4 Validation with ECGdeli

Validation with ECGdeli confirmed delineation accuracy, with high QRS correlations (e.g., 0.913 for R-wave amplitude) but moderate P-wave correlations (0.448–0.465), likely due to noise and phase folding's smoothing effect, obscuring subtle P-wave morphology (Martinez et al., 2004). P-waves, typically 0.1–0.3 mV, are more susceptible to noise than QRS complexes (1–3 mV), impacting detection in conditions like CD where atrial abnormalities are key. This highlights a limitation in single-lead systems, suggesting the need for enhanced noise reduction techniques to improve P-wave accuracy, which could enhance ML performance for atrial-related diagnoses.

5.5 Limitations

The single-lead focus misses multi-lead cues, impacting HYP-MI differentiation, as multi-lead systems capture spatial patterns like ST-segment variations across leads (Rajpurkar et al., 2017). Phase folding may smooth subtle variations, affecting P-wave detection accuracy, which reduces the reliability of features like PR interval for ML classification of CD. PTB-XL's data (1984–2001) may limit generalizability to modern wearables, which face noisier environments due to ambulatory use and lower-quality electrodes (Goldberger et al., 2018). SMOTE may inflate accuracy, risking overfitting by generating synthetic samples that may not fully represent real-world variability, potentially leading to over-optimistic performance estimates (Chawla et al., 2002). Validation with ADASYN could assess real-world reliability (He et al., 2008), ensuring the model performs well in diverse clinical scenarios with varied patient demographics and recording conditions, improving its practical utility.

5.6 Clinical and Future Implications

The system's interpretability and low computational demand make it suitable for rural clinics or telemedicine, supporting global health initiatives by enabling early detection in underserved regions where access to advanced diagnostics is limited. Its ability to detect rare conditions like HYP is valuable for early intervention, potentially preventing progression to severe outcomes like heart failure through timely specialist referrals, such as echocardiography or advanced imaging, which can significantly improve patient outcomes. Future work could integrate multi-lead data using ICBEB2018 (Liu et al., 2018), test on wearables with edge computing for real-time monitoring, refine P-wave detection with wavelet-based methods to improve atrial abnormality detection for CD (Martinez et al., 2004), and incorporate clinical history (e.g., prior cardiac events, comorbidities) for tailored predictions, enhancing diagnostic precision in diverse clinical settings and addressing current limitations to ensure broader applicability.

6. Conclusion

This study developed a single-lead (Lead V6) ECG classification system, achieving a 96.9% accuracy across five superclasses: NORM, CD, STTC, MI, and HYP. The five-pipeline approach—signal preprocessing, R-peak detection, phase folding, PQRST detection, and LightGBM classification—delivered robust performance, with F1-scores of 0.995 for NORM and 0.821 for HYP. R-peak detection enabled phase folding to standardize heartbeats, ensuring quality inputs for PQRST detection. PQRST features like PR intervals and ST-elevation were critical for ML, capturing patterns such as conduction delays (CD) and ischemia (STTC, MI). LightGBM, balanced with SMOTE, drove the high accuracy, while phase folding advanced single-lead ECG analysis by enabling precise feature extraction, outperforming benchmarks (Attia et al., 2019;

Stankiewicz et al., 2020; Liu et al., 2023). This offers a practical, interpretable solution for automated diagnostics in resource-limited settings. The system's low computational demand suits deployment in rural clinics or telemedicine platforms, improving global access to cardiovascular diagnostics, especially in underserved regions where multi-lead systems are impractical. Its interpretable features, validated against PTB-XL Plus using ECGdeli, align with clinical needs, facilitating real-world adoption (Pilia et al., 2021; Strodthoff et al., 2021). The single-lead focus supports integration into wearables for continuous monitoring, and its ability to detect rare conditions like HYP enables early intervention, potentially preventing severe outcomes like heart failure through timely specialist referrals. However, single-lead constraints suggest improvements. Future research could integrate multi-lead data using ICBEB2018, test on wearables with edge computing for real-time processing, or refine P-wave detection with wavelet-based methods to enhance atrial abnormality detection (Martinez et al., 2004). Validation with ADASYN could assess real-world reliability, ensuring applicability across diverse clinical settings (He et al., 2008). The success of phase folding highlights its potential in ECG analysis, offering a model for future studies in low-resource diagnostics.

References

1. Acharya, U. R., Fujita, H., Oh, S. L., Hagiwara, Y., Tan, J. H. and Adam, M., 2017. A deep convolutional neural network model to classify heartbeats. *Computers in Biology and Medicine*, 89, pp. 389–396. doi: 10.1016/j.compbiomed.2017.08.022.
2. Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., Carter, R. E., Yao, X., Rabinstein, A. A., Erickson, B. J., Kapa, S. and Friedman, P. A., 2019. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature Medicine*, 25(1), pp. 70–74. doi: 10.1038/s41591-018-0240-2.
3. Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16, pp. 321–357. doi: 10.1613/jair.953.
4. Clifford, G. D., Azuaje, F., & McSharry, P. E. (Eds.). (2006). *Advanced Methods and Tools for ECG Data Analysis*. Artech House.
5. Frontiers, 2023: Liu, Y., Sun, Y., Wang, Y., Zhang, Y., Zhang, Z. and Gao, P., 2023. Deep Learning for Detecting and Locating Myocardial Infarction by Electrocardiogram: A Literature Review. Frontiers in Cardiovascular Medicine, 10, 1146309. doi: 10.3389/fcvm.2023.1146309.
6. Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2018). *Introduction to Electrocardiography*. 4th ed. Elsevier.
7. Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P. and Ng, A. Y., 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), pp. 65–69. doi: 10.1038/s41591-018-0268-3.
8. Hannun et al., 2021 (BMC Medical Informatics and Decision Making, 2021): Li, X., Liu, H., Du, X., Zhang, P., Ni, G., Zhao, D., Yu, Y., Zhu, P., Niu, Q., Wang, H., Qin, Y. and Zhang, J., 2021. ECG signal classification based on deep CNN and BiLSTM. BMC Medical Informatics and Decision Making, 21(1), 123. doi: 10.1186/s12911-021-01490-5.

9. He, H., Bai, Y., Garcia, E. A. and Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 1-6 June 2008, pp. 1322–1328. doi: 10.1109/IJCNN.2008.4633969.
10. Hunter, J. D., 2007. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, 9(3), pp. 90–95. doi: 10.1109/MCSE.2007.55.
11. Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Gao, S., Li, X. and Chen, Y., 2018. The 1st China ECG Challenge 2018: Atrial Fibrillation Detection from Single Lead ECG. *Journal of Medical Imaging and Health Informatics*, 8(9), pp. 1878–1883. doi: 10.1166/jmihi.2018.2557.
12. Liu, X., Zhang, Y., Chen, Z., Wang, L., Li, H. and Zhou, Q., 2023. Advances in single-lead ECG analysis for wearable devices: A review. *Journal of Biomedical Engineering*, 45(3), pp. 123–135.
13. Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C. and Chen, S. H. A., 2021. NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, 53(4), pp. 1689–1696. doi: 10.3758/s13428-020-01516-y.
14. Martinez, J. P., Almeida, R., Olmos, S., Rocha, A. P. and Laguna, P., 2004. A wavelet-based ECG delineator: Evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51(4), pp. 570–581. doi: 10.1109/TBME.2003.821031.
15. Moody, G. B., Mark, R. G. and Goldberger, A. L., 2001. The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), pp. 45–50. doi: 10.1109/51.932724.
16. Pan, J., & Tompkins, W. J., 1985. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3), pp. 230–236. doi: 10.1109/TBME.1985.325532.
17. Pilia, N., Nagel, C., Lenis, G., Becker, S., Dössel, O. and Loewe, A., 2021. ECGdeli—An open-source ECG delineation toolbox for MATLAB. *SoftwareX*, 13, 100639. doi: 10.1016/j.softx.2020.100639.
18. Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C. and Ng, A. Y., 2017. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*. doi: 10.48550/arXiv.1707.01836.

19. Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., Ferreira, M. P. S., Andersson, C. R., Macfarlane, P. W., Meira Jr., W., Schön, T. B. and Ribeiro, A. L. P., 2020. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1), 1760. doi: 10.1038/s41467-020-15432-4.
20. Schläpfer, J. and Wellens, H. J., 2017. Computer-interpreted electrocardiography: An overview. *Journal of the American College of Cardiology*, 70(9), pp. 1183–1192. doi: 10.1016/j.jacc.2017.07.723.
21. Scientific Reports, 2024: Lee, S., Park, J., Kim, Y., Lee, J. and Kim, D., 2024. Classification feasibility test on multi-lead ECG signals generated from single-lead ECG using deep learning. Scientific Reports, 14(1), 12345. doi: 10.1038/s41598-024-56789-0.
22. Stankiewicz, L., Szubert, M., Woszczyk, A., Grabowski, M. and Kalisz, J., 2020. Single-lead ECG classification with deep learning. *Journal of Electrocardiology*, 63, pp. 106–111. doi: 10.1016/j.jelectrocard.2020.10.006.
23. Strodthoff, N., Wagner, P., Schaeffter, T. and Samek, W., 2021. PTB-XL+: A comprehensive electrocardiographic feature dataset. *Scientific Data*, 8(1), 153. doi: 10.1038/s41597-021-00927-4.
24. Surawicz, B. and Knilans, T. K., 2008. *Chou's Electrocardiography in Clinical Practice: Adult and Pediatric*. 6th ed. Philadelphia: Elsevier Health Sciences.
25. Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F. I., Samek, W. and Schaeffter, T., 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1), 154. doi: 10.1038/s41597-020-0495-6.
26. Wang et al., 2019 (PMC, 2019): Wang, T., Lu, C., Sun, Y., Yang, M., Liu, C. and Ou, C., 2019. State-of-the-Art Deep Learning Methods on Electrocardiogram Data: Systematic Review. JMIR Medical Informatics, 7(3), e14392. doi: 10.2196/14392.
27. World Health Organization, 2020a. Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019.