EEBM 7V87: ECG Beat Classification using SVM

Lila Rana

***Abstract*—In Assignment III, an Electrocardiogram(ECG) beat classification pipeline that leverages Support Vector Machines (SVM) for distinguishing normal and abnormal heartbeats using the MIT-BIH Arrhythmia Database. ECG signals from five records were processed using the Pan-Tompkins algorithm to detect QRS complexes, followed by the extraction of morphological, temporal, and spectral features. Two feature configurations, baseline (8 features) and extended (12 features), were evaluated. The SVM model, trained with an RBF kernel, achieved high classification accuracy, with improved sensitivity and specificity using the extended feature set. Visualizations, including confusion matrices and PCA-based feature separation, validate the effectiveness of the approach for arrhythmia detection in real-world ECG monitoring scenarios.**

***Index Terms*—ECG, SVM, Pan-Tompkin**

I. Introduction

THE ECG signals are essential for non-invasive diagnosis of heart rhythm abnormalities. The MIT-BIH Arrhythmia Database is a widely used resource in biomedical signal processing, providing annotated ECG recordings sampled at 360 Hz. The ECG signals were obtained from PhysioNet's MIT-BIH database [1] using the automated annotation tool [2].

The experimental evaluation is conducted on five selected records (100, 105, 106, 209, and 220) from the MIT-BIH dataset. The proposed approach demonstrates the effectiveness of handcrafted ECG features for arrhythmia detection. The detailed distribution of beat types for each record is shown in Table I. The number of abnormal beat types is less than the normal beat types for each record, and the classes of the dataset are imbalanced.

TABLE I: Distribution of Beat Types for each record

Record	Beat Type (Abnormal, Normal)
100	33 A, 2239 Normal
105	41 V, 2526 Normal
106	520 V, 1507 Normal
209	383 A, 2621 Normal
220	94 A, 1954 Normal

In this study, we used a supervised classification algorithm to classify normal and abnormal heartbeats using SVM. Specifically, we use the Pan-Tompkins algorithm to detect QRS complexes and extract beat-level morphological features, including temporal durations, RR intervals, and segment areas. An SVM classifier is then trained to distinguish between normal (N) and abnormal (A, V) beats based on these features. The methodology used for this work is explained in Section II, and the result is discussed in Section III.

II. Methodology

The ECG classification pipeline involves several sequential steps: data selection, signal preprocessing using the Pan-Tompkins algorithm, morphological feature extraction, and classification using a Support Vector Machine (SVM)[3]. The complete methodology is shown in Figure 1 and described below.

A. Dataset and Annotation Conversion

We utilized five ECG records from the MIT-BIH Arrhythmia Database: 100, 105, 106, 209, and 220. The original annotations were parsed from .txt files [2] and converted to .csv format, storing the sample number and corresponding beat type label for each heartbeat. These serve as ground truth for supervised classification.

B. Preprocessing: Pan-Tompkins QRS Detection

Each ECG signal was sampled at 360 Hz. To enhance the QRS complex for reliable feature extraction, we applied the Pan-Tompkins filtering pipeline, a well-established method for real-time QRS detection in ECG signals. The pipeline consists of the following steps:

1) Normalization: The ECG signal is first mean-subtracted and then amplitude-normalized to reduce baseline variability and bring the signal into a uniform range.

2) Bandpass Filtering: A low-pass and high-pass filter combination is used to suppress baseline drift and high-frequency noise, preserving the key morphological characteristics of the ECG waveform.

3) Derivative and Squaring: The filtered signal is passed through a derivative filter to accentuate slope changes, followed by squaring to emphasize large values and make all data positive.

4) Moving Window Integration (MWI): A moving average filter is applied to smooth the squared signal. This enhances the envelope of the QRS complex, making candidate regions easier to detect through adaptive thresholding.

C. Morphological Feature Extraction

From each detected heartbeat, either an 8-dimensional or 12-dimensional feature vector is extracted depending on the configuration. These features capture temporal, spectral, and morphological characteristics of the ECG waveform. They are categorized as follows:

- **QS width:** Duration between the Q and S peaks within a QRS complex.
- **QR width:** Duration between the Q and R peaks.
- **RS width:** Duration between the R and S peaks.

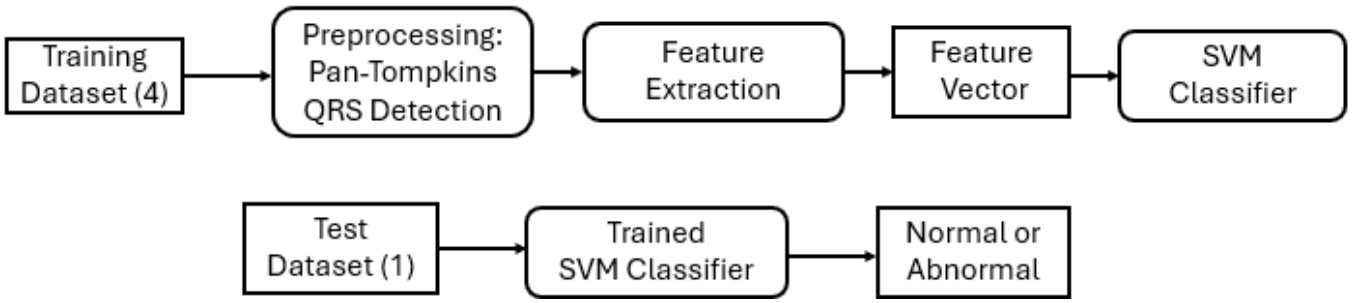

Fig. 1: SVM beat classification model architecture

- **Pre-RR interval:** Time interval between the current R-peak and the previous one.
- **Post-RR interval:** Time interval between the current R-peak and the next one.
- **Mean Power Spectral Density (MPSD):** Computed over the segment from P-wave to T-wave using the squared magnitude of the FFT, indicating the average frequency content of the beat.
- **Area under QR segment:** Estimated using the trapezoidal integration method (`trapz`) applied to the ECG signal between the Q and R peaks.
- **Area under RS segment:** Estimated using `trapz` over the signal segment between R and S peaks.
- **Autocorrelation Value:** Maximum autocorrelation coefficient (excluding the trivial peak) of the ECG segment between Q and S peaks.
- **ST Segment Deviation:** Average amplitude of the ECG segment following the S-peak up to the T-wave, capturing ST elevation or depression.
- **ST Slope:** Linear slope of the ST segment, computed via polynomial fitting.
- **Correlation with Normal Template:** Pearson correlation coefficient between the current beat and a learned average normal beat segment.

The Q, R, and S peaks are identified within regions highlighted by the MWI-enhanced signal. The P and T waves are detected relative to the QRS complex, based on local maxima within appropriate temporal windows. Edge-truncated or incomplete beats are discarded to ensure the reliability of extracted features. The extracted ECG segment with the peaks P, Q, R, S, and T is shown in Figure 2.

D. SVM Model

An SVM model is employed to distinguish between normal and abnormal heartbeats based on the extracted morphological features. SVM is a supervised learning algorithm known for its effectiveness in binary classification tasks, particularly when the decision boundary is nonlinear.

Training : All extracted feature vectors and their corresponding labels were aggregated from the selected ECG records. To ensure balanced representation of classes, a stratified 70/30 split was applied, preserving the proportion of normal and abnormal beats across training and testing sets, which is shown in Table II.

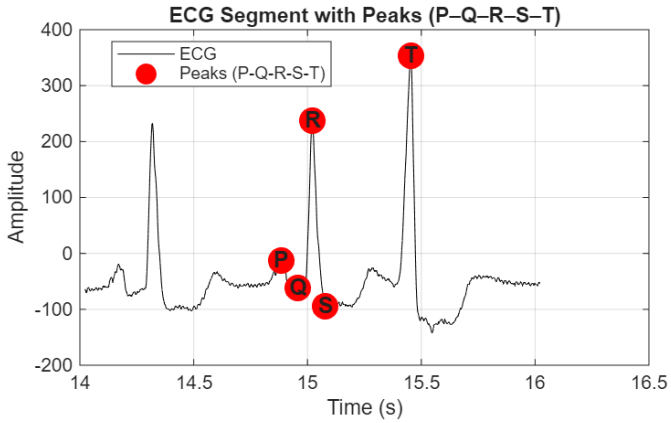

Fig. 2: ECG waveform showing P, Q, R, S, and T peaks.

Model Configuration: The SVM classifier was configured with the following parameters:

- **Kernel:** Radial Basis Function (RBF), suitable for capturing non-linear decision boundaries.
- **Kernel Scale:** Automatically selected using heuristic methods.
- **Box Constraint:** Set to infinity (`Inf`) to enforce a hard margin and minimize misclassification.
- **Standardization:** All features were standardized to zero mean and unit variance before training.
- **Solver:** Iterative Single Data Algorithm (ISDA), chosen for its efficiency on medium-sized datasets.

Classification and Prediction: The trained SVM model was used to classify each test beat as normal (label `1`) or abnormal (label `0`). Classification results were evaluated using metrics such as accuracy, false positive/negative rates, and beat counts.

Label Assignment: The reference labels were derived from the annotation files of the MIT-BIH dataset. For each detected beat, the nearest annotated R-peak was matched, and its label was assigned accordingly.

TABLE II: Dataset Statistics

Total Beats	Normal Beats	Abnormal Beats
11,998	10,978 (7685 train, 3293 test)	1,020 (714 train, 306 test)

Principal Component Analysis (PCA) was applied to project the extracted features into a two-dimensional space for visualization, which is shown in Figure 3. The resulting distribution

revealed that normal and abnormal beats tend to form distinct, non-overlapping clusters. This separation in PCA space validates the discriminative power of the extracted features and supports the suitability of SVM as the classification model.

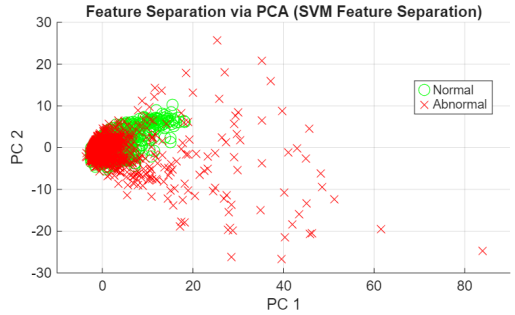

Fig. 3: Feature separation using PCA for SVM input features. Normal and abnormal beats show distinct clusters, validating feature effectiveness.

III. Results and Discussion

We evaluated the performance of our ECG classification using two different feature configurations: a baseline set of 8 features and an extended set of 12 features. The classification was performed using an SVM, and the results are shown in the confusion matrices below.

A. Feature Set Comparison

The 8-feature configuration includes:

- QS width, QR width, RS width
- Pre-RR interval, Post-RR interval
- Mean Power Spectral Density (MPSD)
- Area under QR segment
- Area under RS segment

The extended 12-feature configuration additionally includes:

- Autocorrelation value
- ST segment deviation
- Slope of ST segment
- Correlation with normal template

B. Classification Metrics

TABLE III: Classification Metrics for 8 vs. 12 Feature Configurations

Metric	8 Features	12 Features
True Positives (Normal → Normal)	98.97%	99.30%
True Negatives (Abnormal → Abnormal)	90.52%	93.46%
False Negatives (Normal → Abnormal)	1.03%	0.70%
False Positives (Abnormal → Normal)	9.48%	6.54%
Accuracy	98.25%	98.81%

As shown in Table III, incorporating the additional four features improved the model's sensitivity and specificity, reducing both false negatives and false positives.

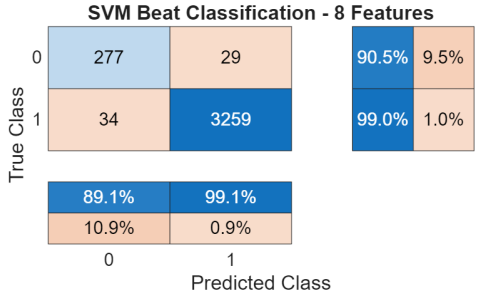

Fig. 4: SVM Beat classification using 8 features.

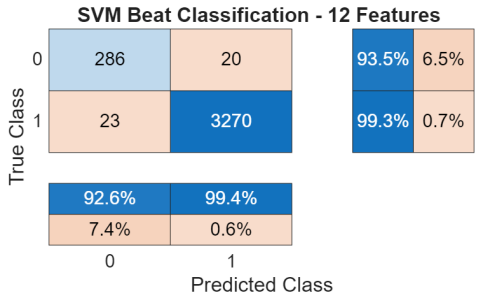

Fig. 5: SVM Beat classification using 12 features.

C. Confusion Matrices

The respective confusion matrices for SVM beat classification using 8 features and 12 features are shown in Figure 4 and Figure 5. To further illustrate classification performance, Figure 6 shows an ECG waveform with beat-level annotations based on SVM predictions. The signal corresponds to a 20-second segment from Record 105 of the MIT-BIH Arrhythmia Database. Normal beats, correctly identified by the classifier, are marked with green circles, while an abnormal beat is highlighted in a red circle. This visualization demonstrates the classifier's ability to isolate abnormal beats amid a majority of normal ones, reflecting its practical utility in detecting rare arrhythmias during continuous monitoring.

One of the key challenges encountered was the class imbalance inherent in the dataset. The number of abnormal beats, particularly in records 100, 105, and 220, is significantly lower than the number of normal beats. This imbalance affected the classifier's ability to learn from underrepresented classes. To mitigate this, a stratified split was applied within each

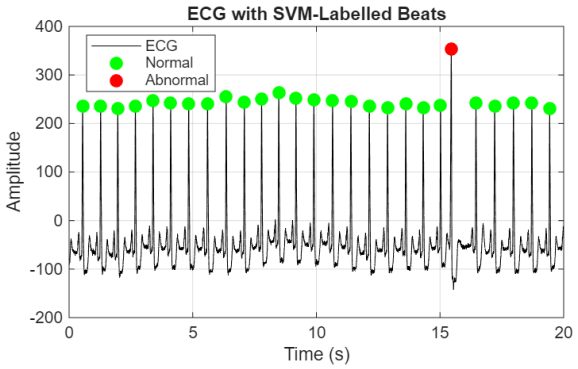

Fig. 6: ECG waveform with SVM-labeled beats

record, ensuring that both normal and abnormal beats were proportionally represented in training and testing sets. By applying this record-wise stratification, the classifier maintained consistent performance across different subjects and recording conditions, thus improving generalizability and robustness.

ACKNOWLEDGMENT

I want to offer thanks to Prof. Dr. Lakshman Tamil for offering this course, EEBM 7V87 Special Topics in Biomedical Application of EE to enrich the understanding of Telemedicine and AI for healthcare.

REFERENCES

[1] Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., Ivanov, P. Ch., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (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

[2] PhysioNet ATM: PhysioBank Archives Index. https://archive.physionet.org/cgi-bin/atm/ATM

[3] Bulusu, S. C. (2010). *Detection of ECG transient ST-segment episodes and machine learning-based heartbeat classification* (Master's thesis). The University of Texas at Dallas.