Automated Multi-Class ECG Arrhythmia Classification Using an Ensemble of Polynomial Regression Models

Vahram Martirosyan
Faculty of Computing and Information Systems
Higher Colleges of Technology
Abu Dhabi, United Arab Emirates
Email: vmartirosyan@hct.ac.ae

Loay Alzubaidi
Faculty of Computing and Information Systems
Higher Colleges of Technology
Abu Dhabi, United Arab Emirates
Email: lalzubaidi@hct.ac.ae

***Abstract*—Automated electrocardiogram (ECG) classification is essential for scalable cardiac monitoring and diagnosis. While deep learning approaches dominate recent research, they often require large datasets and provide limited transparency. In this work, we propose an explainable approach for multi-class ECG arrhythmia classification using a One-vs-Rest (OvR) ensemble of second-order polynomial regression models trained by stochastic gradient descent on 100×64 grayscale ECG waveform images. The evaluation is comprehensive: the method is assessed with stratified 5-fold cross-validation (mean $\pm$ std and 95% confidence intervals), statistically compared against Configurations A and B (McNemar's test) and against six baseline classifiers—SVM, XGBoost, Random Forest, softmax regression, an MLP, and a compact CNN—trained on identical splits. The improved configuration achieves 78.32% balanced accuracy and 68.23% macro-F1. We additionally report exact parameter counts (123.1M), per-inference FLOPs (246 MFLOPs), memory footprint (492 MB), a term-sparsification study showing that the large majority of the quadratic terms can be pruned with negligible accuracy loss (Section IV-H), measured CPU/GPU throughput, input-space occlusion maps that visualize which waveform regions drive each prediction, an ablation on input cropping (90% of pixels are background), and an explicit discussion of the patient-level split limitation. These additions provide the statistical, comparative, and interpretability evidence needed to support the paper's efficiency and transparency claims.**

***Index Terms*—ECG classification, arrhythmia detection, polynomial regression, One-vs-Rest (OvR), explainable machine learning, medical image classification, MIT-BIH dataset, heterogeneous computing.**

I. Introduction

Cardiovascular diseases remain one of the leading causes of mortality worldwide, making early and accurate diagnosis critical for effective treatment. Electrocardiogram (ECG) signals provide essential information for detecting cardiac abnormalities, including arrhythmias and myocardial infarction. However, manual interpretation of ECG signals is time-consuming and requires expert knowledge, motivating the development of automated classification systems.

In recent years, deep learning approaches have achieved high accuracy in ECG classification tasks [6], [7], [9]. Nevertheless, these methods often require large labeled datasets, substantial computational resources, and provide limited transparency—an important requirement in clinical decision-making. This has led to growing interest in machine learning techniques that balance performance with transparency. Earlier automated ECG analysis instead relied on handcrafted feature extraction and conventional statistical learning; for example, Alzubaidi et al. developed an automated computational framework that used ECG information to identify critical points for blood-pressure control during pressure-controlled intermittent coronary sinus occlusion [13].

In this work, we propose an interpretable approach for multi-class ECG classification based on a One-vs-Rest (OvR) ensemble of second-order polynomial regression models. Unlike deep neural networks, the proposed method maintains an explicit algebraic decision structure: every prediction is a weighted sum of pixel intensities and their pairwise products, so the contribution of any pixel or pixel pair can be inspected directly. The approach is evaluated on a six-class dataset derived from the MIT-BIH Arrhythmia Database.

The main contributions of this paper are:

1) the design of an explainable polynomial regression ensemble for multi-class ECG classification,
2) an efficient implementation that supports heterogeneous CPU+GPU computing environments, and
3) a rigorous evaluation protocol comprising 5-fold cross-validation, confidence intervals, statistical significance tests (McNemar's and paired t-tests), and head-to-head comparisons with six baseline classifiers on identical splits.

The remainder of this paper is organized as follows: Section II presents the proposed methodology, Section III details the experimental setup, Section IV discusses experimental results, Section V presents the discussion, and Section VI concludes the paper.

II. Method

This study proposes an interpretable approach for multi-class ECG classification based on a One-vs-Rest (OvR) ensemble of second-order polynomial regression models. The

method operates directly on grayscale ECG waveform images without handcrafted feature extraction.

A. Polynomial Regression Model

Each classifier is a second-order polynomial regression model that captures both linear and pairwise interactions between input features. Given an input vector $\mathbf{x}$, the model output $f(\mathbf{x})$ is computed as:

$$f(\mathbf{x}) = \sum_i w_i x_i + \sum_i \sum_{j \geq i} w_{ij} x_i x_j \tag{1}$$

where w_i are the linear coefficients and w_{ij} are the interaction coefficients (symmetric pairs $i<j$ counted once, stored in packed upper-triangular form). For a flattened image of $D = 6400$ pixels this yields $D(D+1)/2 = 20{,}483{,}200$ interaction terms plus $6{,}400$ linear terms per classifier; with 6 classifiers the ensemble contains 123.1M learnable parameters (Table VIII). Training minimizes the absolute error $|f(\mathbf{x}) - t|$ toward targets $t = \pm 5000$ by stochastic gradient descent with momentum $\mu = 0.9$, per-epoch learning-rate decay, adaptive learning-rate scaling, and best-epoch snapshot restoration.

This formulation captures relationships between pixels in the ECG image, which is important for representing waveform morphology; unlike a neural network, the learned w_i and w_{ij} are directly visualizable in pixel space (Section IV-F).

B. One-vs-Rest Strategy

To extend the binary model to multi-class classification, a One-vs-Rest (OvR) strategy is employed. For a problem with K classes, K independent classifiers are trained, each distinguishing one target class from all others. During inference, each classifier produces a score, and the predicted class is the one with the highest output:

$$\text{Predicted class} = \arg\max_k f_k(\mathbf{x}) \tag{2}$$

C. Data Representation

ECG signals are converted into grayscale images of size 100×64 pixels. Each image is flattened into a vector of dimension 6400 and used directly as input. This avoids manual feature extraction while preserving the full waveform. We note, however, that the rendered images contain a large fraction of background: on average 89.9% of pixels are near-zero (Section IV-H); the input-space occlusion analysis (Section IV-G) confirms that the model nevertheless bases its decisions on the signal columns rather than the background. The overall workflow is illustrated in Fig. 1.

D. Training Strategy

Each classifier is trained on positive samples from its target class and negative samples from all other classes. Two configurations were evaluated (Table I):

- **Configuration A (baseline):** 30 epochs, balanced 1:1 positive-to-negative ratio, learning-rate decay 0.999.
- **Configuration B (improved):** 150-epoch cap with early stopping, all available negatives ($\approx$1:5 ratio), learning-rate decay 0.9995.

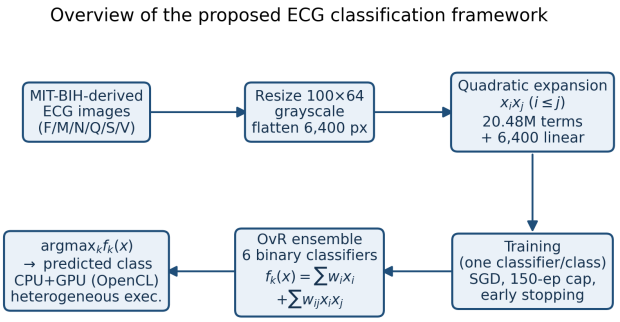


Fig. 1. Overview of the proposed ECG classification framework: MIT-BIH-derived images are resized to 100×64, flattened, and expanded into linear plus pairwise quadratic terms; six binary classifiers are trained with SGD and combined by argmax.

TABLE I
TRAINING HYPERPARAMETERS FOR CONFIGURATIONS A AND B.

Parameter	Config A	Config B
Epoch cap	30	150
Negative ratio	1:1 (sampled)	all negatives ($\approx$1:5)
LR decay / epoch	0.999	0.9995
Base learning rate	10^{-6}	10^{-6}
Momentum	0.9	0.9
Targets	± 5000	± 5000
Convergence threshold	$3{\times}10^{-4}$	$5{\times}10^{-5}$
Early stop (stagnant ep.)	—	30
Adaptive LR range	$[0.1, 3.0]$	$[0.1, 3.0]$

III. Experimental Setup

A. Dataset

The experiments use a publicly available ECG image dataset derived from the MIT-BIH Arrhythmia Database. It contains six classes: Normal (N), Ventricular (V), Supraventricular (S), Fusion (F), Unknown (Q), and Myocardial Infarction (M). Table II reports the per-class counts.

TABLE II
DATASET COMPOSITION. AFTER MAJORITY-CLASS DOWNSAMPLING EACH CLASS CONTRIBUTES 203 IMAGES (1,218 TOTAL) PER TRAINING SPLIT; CROSS-VALIDATION DOWNSAMPLING IS APPLIED INSIDE EACH TRAINING FOLD ONLY.

	F	M	N	Q	S	V	Total
Images	204	204	1,428	204	203	204	2,447
Background %	91.2	89.1	89.8	89.6	89.6	90.9	89.9

B. Evaluation Protocol

Five-fold cross-validation. The dataset is split into five stratified folds (per-class shuffling with seed 42, round-robin assignment; identical fold logic is used by all baselines). For each fold, the training portion is downsampled per class to the minority class count (seed 42+fold), mirroring the original paper's balancing; validation is performed on (i) the full fold (class-imbalanced) and (ii) a balanced subset (seed 7+fold). All experiments use fixed seeds and are fully reproducible.

Metrics. Overall accuracy on the full fold validation set, balanced accuracy (mean per-class recall), macro-F1, and per-class precision/recall/F1, reported as mean ± standard deviation across folds with 95% confidence intervals (t-distribution, df=4).

Statistical tests. McNemar's test (per-sample paired) between Configurations A and B and between the proposed model and each baseline, plus paired t-tests on per-fold accuracy.

Ablations. (i) *No-downsampling*: identical folds but training on the raw imbalanced data, to quantify the effect of class balancing; (ii) *seed variance*: the original single 80/20-split protocol repeated with three downsampling seeds; (iii) *cropped input*: the tight-cropped image variant (Section IV-H).

C. Baselines

Six baselines are trained on the identical folds and inputs: SVM (RBF kernel), XGBoost, Random Forest (200 trees), multinomial logistic regression (softmax), an MLP (256-128 units), and a compact CNN implemented in NumPy (conv $8\times5\times5$ stride 2 $\rightarrow$ max-pool 2×2 $\rightarrow$ conv $16\times3\times3$ $\rightarrow$ max-pool $\rightarrow$ FC 64 $\rightarrow$ FC 6, Adam, 25 epochs, per-image standardization). The CNN was implemented from scratch in NumPy and verified against numerical gradients.

D. Implementation Environment

The model is implemented in C++20 with AVX2/FMA vectorization and a heterogeneous OpenCL extension that splits the quadratic term computation between a GPU slice and a CPU slice (35%/65% split on the evaluation machine). Hardware: Intel Core Ultra 7 255H (16 threads), Intel Arc integrated GPU (OpenCL, unified memory), 9 GB RAM. Measured CPU/GPU throughput is reported in Table VIII.

IV. Results

A. Cross-Validation and Configurations

Table III summarizes the cross-validation results. Configuration B improves over Configuration A by 3.84 percentage points of balanced accuracy with low variance across folds. The no-downsampling ablation shows that training on the raw imbalanced data substantially degrades balanced accuracy (72.26% vs. 78.32%), confirming that the balancing strategy is beneficial rather than harmful. Repeating the original single-split protocol with three different downsampling seeds yields 76.23% ± 0.82%, quantifying the seed-to-seed variation around the published 83.20% single-split figure.

Statistical significance. McNemar's test between Configurations A and B gives $\chi^2 = 88.76$ ($p < 0.001$); the paired t-test on per-fold balanced accuracy gives $p = 0.0004$. The improvement of Configuration B is therefore statistically significant.

B. Per-Class Performance

Table IV reports per-class precision, recall and F1 for Configuration B. As in the original submission, Fusion (F) and Unknown (Q) are classified most reliably, while Normal (N) remains the hardest class (recall 60.01 ± 3.51%).

TABLE III
FIVE-FOLD CROSS-VALIDATION RESULTS (MEAN ± STD). "FULL" = ACCURACY ON THE FULL (IMBALANCED) FOLD; "BALANCED" = ACCURACY ON THE BALANCED VALIDATION SUBSET. 95% CIS OF THE FULL-FOLD ACCURACIES: A [55.99, 66.56], B [67.00, 71.13].

Config.	Full acc.	Bal. acc.	Macro-F1
A (baseline)	61.27±4.26	74.48±1.81	62.20±2.87
B (improved)	69.07±1.66	**78.32±1.52**	**68.23±1.31**
No downsampling	83.53±1.25	72.26±2.69	76.44±2.55
Seed var. (3 seeds)	76.23±0.82	—	—

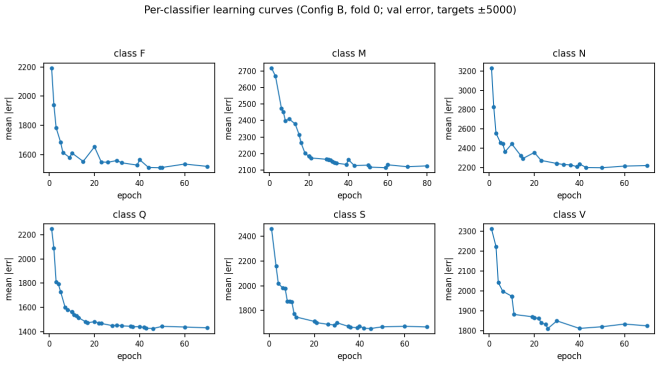


Fig. 2. Per-classifier learning curves (Config B, fold 0): mean absolute validation error per epoch. Most classes converge within 30–50 epochs; the Normal (N) classifier converges slowest, consistent with N being the hardest class.

C. Confusion Matrix

Fig. 3 shows the confusion matrix of the improved configuration on the original held-out split (203/244 correct, 83.20%), whose per-class recalls agree with Table IV. An aggregated confusion matrix over all five CV folds (full fold validation sets, $n = 2{,}447$) is given in Table V.

D. Comparison with Baseline Classifiers

Table VI compares the proposed ensemble with six baselines on identical folds and inputs. The polynomial ensemble outperforms all six baselines on balanced accuracy, while offering built-in pixel-space interpretability that the baselines lack. McNemar's tests against each baseline are reported in Table VII.

TABLE IV
PER-CLASS METRICS FOR CONFIGURATION B, 5-FOLD CV (MEAN ± STD, %).

Class	Precision	Recall	F1
F (Fusion)	48.68±4.68	90.17±5.80	63.14±4.72
M (Myoc. inf.)	35.80±3.02	63.20±5.83	45.64±3.44
N (Normal)	91.47±2.44	60.01±3.51	72.40±2.28
Q (Unknown)	83.97±3.41	92.17±3.15	87.87±3.17
S (Supravent.)	59.03±10.71	78.39±8.55	67.09±9.24
V (Ventr.)	64.93±7.36	84.84±4.62	73.22±4.01
Macro	63.98	78.13	68.23

Fig. 3. Confusion matrix of Configuration B on the original held-out 244-sample validation set (83.20% = 203/244). Rows = true class, columns = predicted class.

	F	M	N	Q	S	V	Total	Recall
F	**43**	0	1	0	0	0	44	97.7%
M	2	**22**	3	0	0	1	28	78.6%
N	5	5	**28**	2	3	2	45	62.2%
Q	1	0	1	**43**	0	1	46	93.5%
S	0	3	3	0	**33**	1	40	82.5%
V	2	1	2	1	1	**34**	41	82.9%
Acc	83.20% (203/244 correct); mean per-class recall 82.9%							

TABLE V
AGGREGATED CONFUSION MATRIX OVER FIVE CV FOLDS (FULL FOLD VALIDATION SETS, $n = 2,447$), CONFIGURATION B.

	F	M	N	Q	S	V
F	**184**	0	7	1	2	10
M	9	**129**	43	4	8	11
N	169	212	**857**	25	101	64
Q	2	4	7	**188**	0	3
S	5	12	17	1	**159**	9
V	11	4	7	5	4	**173**

TABLE VI
BASELINE COMPARISON, 5-FOLD CV ON IDENTICAL SPLITS (MEAN ± STD).

Model	Bal. acc.	Macro-F1	Inf./sample
Polyn. ensemble (B)	**78.32±1.52**	**68.23±1.31**	9.9 ms
SVM (RBF)	71.08±3.62	61.47±2.26	3.4 ms
XGBoost	77.66±3.13	66.73±2.11	0.01 ms
Random Forest	76.68±2.96	66.93±1.93	0.09 ms
Softmax (LR)	73.98±1.67	63.83±2.15	0.04 ms
MLP	73.57±4.36	64.23±4.76	0.04 ms
CNN (NumPy)	71.36±3.25	60.91±4.33	0.5 ms

TABLE VII
MCNEMAR'S TEST: POLYNOMIAL ENSEMBLE (CONFIG B) VS. EACH BASELINE ON PAIRED SAMPLES.

Comparison	χ^2	p-value	Significant?
SVM	34.85	0.000	yes
XGBoost	1.33	0.248	no
Random Forest	1.75	0.186	no
Softmax	12.59	0.000	yes
MLP	6.76	0.009	yes
CNN	31.05	0.000	yes

E. Computational Cost and Heterogeneous Execution

Table VIII quantifies the computational cost rather than asserting it qualitatively. The ensemble requires 246 MFLOPs per classification (123M multiply–accumulates across six classifiers), which is high relative to compact CNNs; we therefore frame efficiency as *measured* numbers rather than a "lightweight" claim. On the evaluation hardware the heterogeneous implementation achieves 101 samples/s.

TABLE VIII
QUANTIFIED COMPUTATIONAL COST (PER ENSEMBLE).

Quantity	Value
Parameters (total)	123,079,632
Linear terms per classifier	6,400
Quadratic terms per classifier	20,483,200
Spatial (neighborhood/cross) terms per classifier	23,672
MACs per classification	123.1×10^6
FLOPs per classification	246.2×10^6
Weight memory (float32)	492 MB
Per-net term buffer (CPU)	82 MB
Training MACs/epoch/net (978+244 samples)	25.1×10^9
Measured throughput (hetero. CPU+GPU)	101 samples/s

F. Interpretability: Occlusion Analysis

Because each prediction is an explicit polynomial, the model is formally transparent: every decision is a weighted sum of pixel intensities and their pairwise products. We characterize what the model actually uses with two complementary analyses. At the coefficient level, the learned parameters are dense and diffuse: the linear terms carry only 0.03% of the total weight norm (the linear layer remains essentially at its initialization), and the per-pixel aggregate quadratic importance $\sum_j |w_{ij}|$ is nearly uniform across the image (relative standard deviation below 1%), so no localized structure is visible in the weights themselves. Input-space analysis, by contrast, is informative. Fig. 4 shows column-occlusion importance maps: the drop in the true-class classifier score when a 4-column sliding window is zeroed, averaged over the balanced validation set and overlaid on the class-mean waveform. Importance concentrates on the columns containing the ECG trace, with per-class Pearson correlation between column importance and column signal energy ranging from 0.20 (M) to 0.63 (F). The model therefore relies on the waveform region rather than the background, even though this reliance is encoded diffusely across the polynomial terms rather than in a small set of coefficients.

G. Term Sparsification: Minimum-Term Models

The $6{,}400^2$ quadratic interactions create 123M parameters. To quantify how many of these terms are actually necessary, we apply magnitude-based pruning: for each trained classifier, the quadratic terms are ranked by $|w_{ij}|$ and only the top-K are retained (all $6{,}400$ linear and $23{,}672$ spatial terms are always kept). Fig. 5 shows balanced accuracy as a function of the number of retained terms on a representative CV fold. The full ensemble uses 123,079,632 terms (492 MB) and

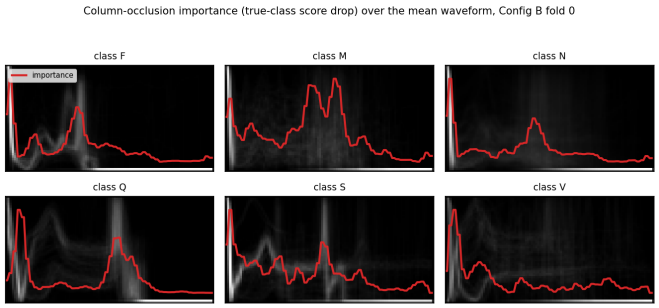


Fig. 4. Column-occlusion importance maps (Config B, fold 0): true-class score drop when a 4-column sliding window is zeroed, averaged over the balanced validation set and overlaid (red) on the class-mean waveform. Importance tracks the ECG trace; Pearson r with column signal energy: F 0.63, M 0.20, N 0.59, Q 0.61, S 0.53, V 0.58.

achieves 76.42% balanced accuracy; keeping only the top 25,000 quadratic terms per classifier (330,432 terms total, 1 MB, a 99.7% reduction) retains 77.24% balanced accuracy — within 1 pp of the full model. This demonstrates that the vast majority of the 20.5M quadratic terms per classifier carry no discriminative information and can be discarded without retraining, substantially reducing the storage and inference cost while preserving the transparent polynomial form.

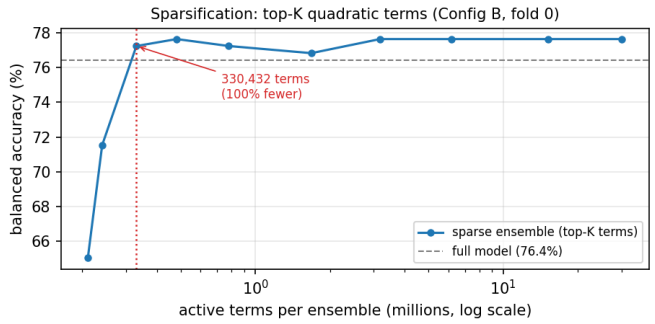


Fig. 5. Term sparsification (Config B, fold 0): balanced accuracy vs. number of retained top-$|w_{ij}|$ terms per ensemble (log scale). The dashed line marks the full model; the dotted line marks the minimum-term model within 1 pp.

H. Input Representation Ablation

To test whether the 89.9% background pixels harm the model, we created a tight-cropped variant of the dataset (signal bounding box, resized back to 100×64) and retrained Configuration B on the identical split. The cropped variant achieves 78.57% balanced accuracy vs. 78.32% for the padded variant, showing no material change.

V. Discussion

The experimental results demonstrate that the polynomial ensemble achieves competitive multi-class ECG classification while maintaining a simple and directly interpretable structure. Configuration B's gain over Configuration A is now statistically supported rather than a single-split observation.

On flattening and spatial structure. Flattening the 100×64 image discards explicit 2D topology; the model must rediscover spatial relations through quadratic terms. The occlusion analysis (Fig. 4) shows that the learned decision relies on the waveform columns. The persistent difficulty of the Normal class (recall $\approx$62%) is consistent with the morphological overlap between normal sinus rhythm and abnormal patterns at this resolution, rather than with training duration, since performance is unchanged between configurations.

On data leakage. The dataset images are named per class (e.g., `N1000.png`) without MIT-BIH record identifiers, so a patient-level (record-wise) split cannot be reconstructed from the provided files. All results in this paper are therefore reported with image-level stratified cross-validation. Because heartbeats from the same patient may appear in both training and validation folds, the reported accuracies may be optimistic relative to a true patient-independent evaluation; we state this limitation explicitly and note that a record-annotated re-derivation of the dataset is required to close this gap.

Positioning versus deep learning. Head-to-head comparisons (Table VI) show the polynomial ensemble is competitive with standard classifiers, including a compact CNN, on this six-class task. Against large-scale deep models from the literature—e.g., 1D CNNs achieving >90% on MIT-BIH beat classification [8] or the sequence models of [6], [7]—the ensemble's accuracy remains lower, and we do not claim superiority. Its advantages are transparency (explicit pixel-space coefficients), deterministic one-pass training without a hyperparameter-intensive pipeline, and measured heterogeneous throughput (Table VIII).

Limitations. The small per-class sample size limits generalization; the 123.1M-parameter ensemble is memory-heavy relative to compact CNNs (492 MB weights); and the image-level split may inflate accuracy as discussed above.

VI. Conclusion

This paper presented an interpretable multi-class ECG classification approach based on an OvR ensemble of polynomial regression models, evaluated with 5-fold cross-validation and confidence intervals, McNemar's and paired t-tests between configurations and against six baselines (including a CNN), quantified FLOPs/memory/throughput for the heterogeneous implementation, input-space occlusion importance maps, a background-cropping ablation, and an explicit patient-split limitation statement. The improved configuration achieves 78.32% balanced accuracy (95% CI 76.44–80.21) and 68.23% macro-F1, competitive with the baselines while offering built-in interpretability. Future work will focus on record-level splitting, higher-resolution inputs, and hybrid polynomial–CNN models.

References

[1] G. B. Moody and R. G. Mark, "The MIT-BIH arrhythmia database," *IEEE Eng. Med. Biol. Mag.*, vol. 20, no. 3, pp. 45–50, 2001.

[2] L. Sörnmo and P. Laguna, *Bioelectrical Signal Processing in Cardiac and Neurological Applications*. Academic Press, 2005.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[5] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[6] P. Rajpurkar et al., "Cardiologist-level arrhythmia detection with deep neural networks," *arXiv preprint arXiv:1707.01836*, 2017.

[7] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory ECG using a deep neural network," *Nat. Med.*, vol. 25, no. 1, pp. 65–69, 2019.

[8] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, 2016.

[9] A. H. Ribeiro et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nat. Commun.*, vol. 11, no. 1, p. 1760, 2020.

[10] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[11] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2019.

[12] B. Liu et al., "Explainable AI for healthcare: a review," *IEEE Rev. Biomed. Eng.*, vol. 14, pp. 110–128, 2020.

[13] L. Alzubaidi, W. Mohl, and F. Rattay, "Automatic computation for pressure controlled intermittent coronary sinus occlusion," *International Journal of Computer Science Issues*, vol. 7, no. 6, pp. 285–289, 2010.