University of Science and Technology of Hanoi

Classification Report

Machine Learning and Data Mining

Le Viet Hoang Lam – 22BI13235

TABLE OF CONTENT

I. Logistic Regression

1. Analyze the dataset

The MIT-BIH Arrhythmia heartbeat dataset is organized as fixed-length ECG segments. Each sample contains 187 amplitude values and one categorical label in $\{0, 1, 2, 3, 4\}$. The training set consists of 87,554 samples, while the test set contains 21,892 samples.

Class imbalance. The dataset exhibits severe imbalance. The number of samples in each class in the training set is: Class 0: 72,471; Class 1: 2,223; Class 2: 5,788; Class 3: 641; Class 4: 6,431. This distribution is illustrated in Figure 1.

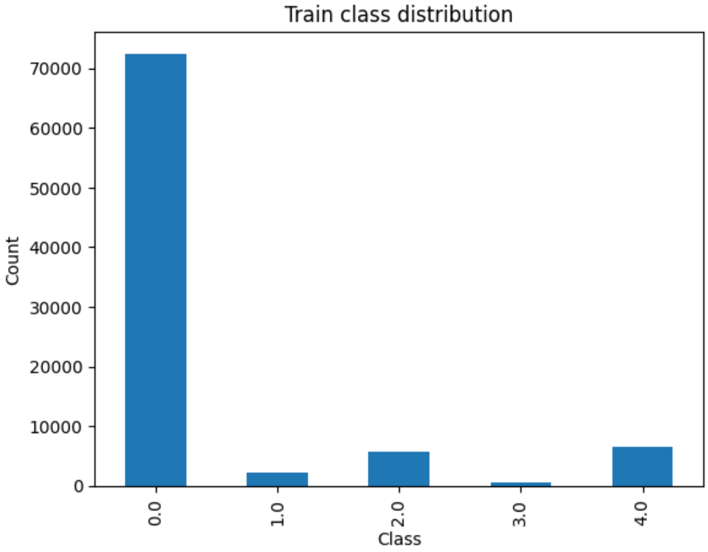

Figure 1: Class distribution of ECG heartbeat training data.

Data quality. No missing values are detected in both training and test sets. Therefore, the dataset requires no imputation or extensive cleaning.

Signal characteristics. Representative ECG waveforms from each class reveal non-linear temporal patterns and significant intra-class variability. Such complexity suggests that linear separability in the raw feature space is limited.

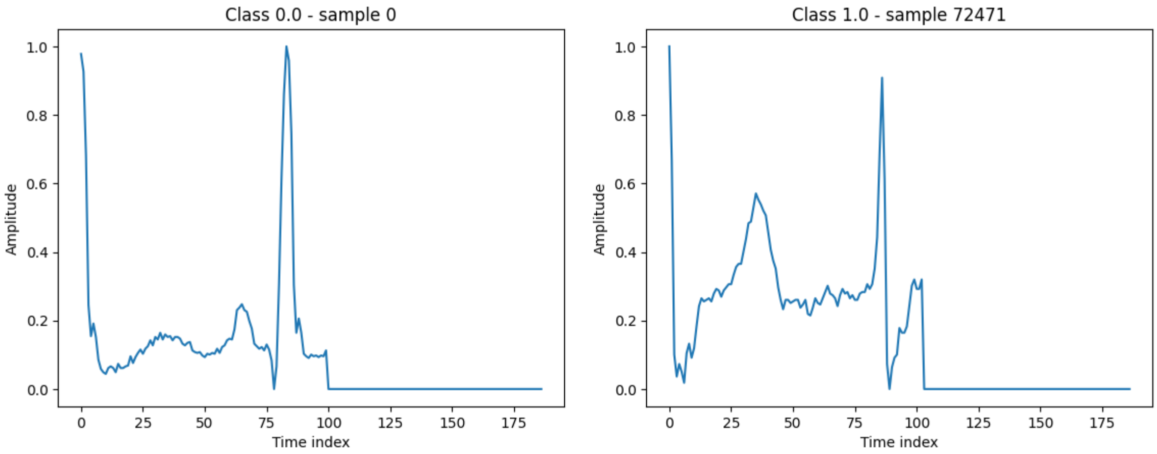

Figure 2: Sample ECG signals from Class 0 and Class 1.

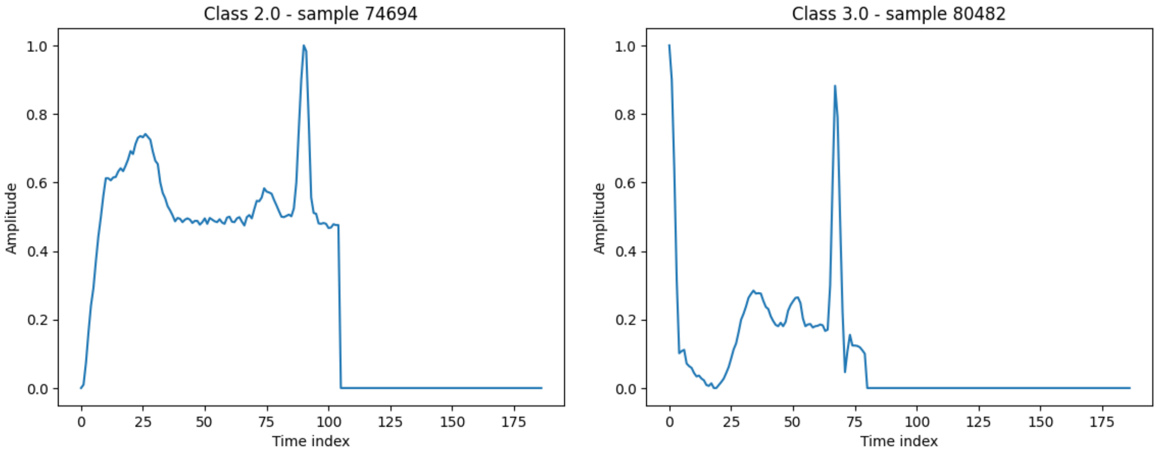

Figure 3: Sample ECG signals from Class 2 and Class 3.

2. Logistic Regression model

The dataset is split into training and validation sets using an 80–20 stratified strategy to preserve class proportions.

All features are standardized using `StandardScaler` to ensure stable optimization.

The Logistic Regression classifier models posterior probabilities using the softmax function:

$$P(y=k|x) = \frac{\exp(w_k^\top x + b_k)}{\sum_j \exp(w_j^\top x + b_j)}$$

The model is trained using the `lbfgs` solver with a maximum of 3000 iterations.

This linear model serves as a computationally efficient baseline for ECG classification.

Table 1: Hyperparameters of the Logistic Regression classifier

Parameter	Value
Solver	lbfgs
Maximum iterations	3000
Regularization strength (C)	1.0
Penalty	Default (L2)
Class weight	None
Tolerance	1×10^{-4}
Fit intercept	True
Warm start	False

The `lbfgs` solver is a quasi-Newton optimization method suitable for multi-class logistic regression and large datasets. A high iteration limit (3000) ensures convergence in the high-dimensional feature space. The regularization parameter $C = 1.0$ corresponds to moderate L2 regularization, balancing bias and variance.

No class weighting is applied in the baseline model, causing the loss function to be dominated by the majority class. This design choice explains the high overall accuracy but reduced recall for minority heartbeat categories observed in the confusion matrices.

3. Error calculation

Due to class imbalance, performance is evaluated using per-class metrics and confusion matrices in addition to overall accuracy.

Table 2: Validation classification metrics.

Class	Precision	Recall	F1-score
0	0.925	0.996	0.959
1	0.823	0.452	0.583
2	0.633	0.336	0.439
3	0.585	0.297	0.395
4	0.936	0.890	0.913

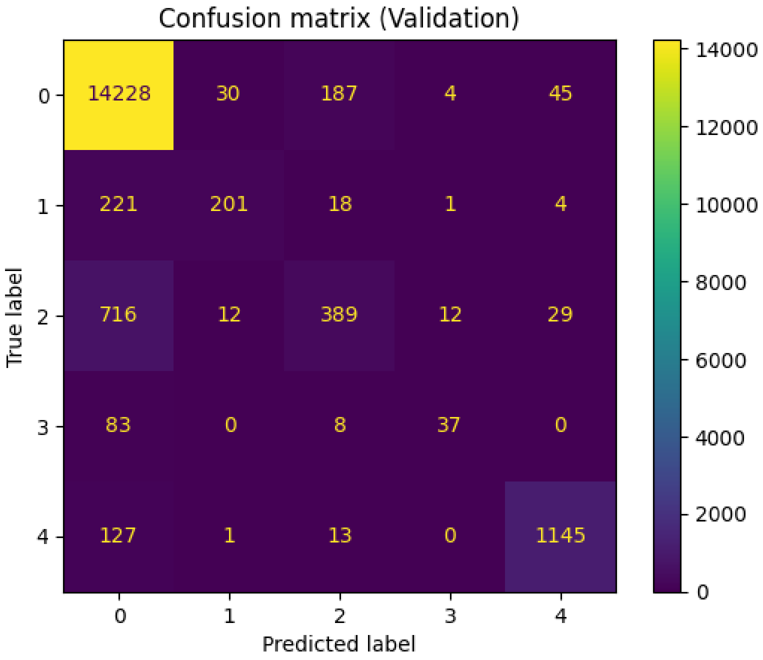

Figure 4: Confusion matrix on validation set.

Validation performance. The validation accuracy is approximately 91.7%. However, minority classes exhibit substantially lower recall compared to majority classes.

Table 3: Test classification metrics.

Class	Precision	Recall	F1-score
0	0.925	0.995	0.959
1	0.882	0.419	0.568
2	0.667	0.335	0.447
3	0.582	0.375	0.455
4	0.957	0.888	0.916

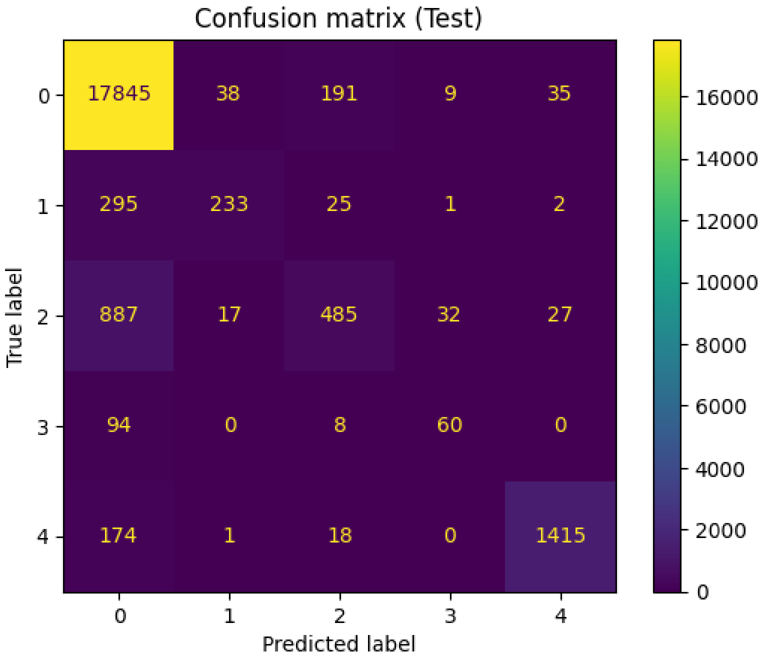

Figure 5: Confusion matrix on test set.

Test performance.

Error interpretation. The confusion matrices show that minority classes (particularly Classes 2 and 3) are frequently misclassified as Class 0. This behavior reflects the combined effects of class imbalance and the linear nature of Logistic Regression.

Limitations and improvements. Without reweighting or resampling, the loss function prioritizes majority-class performance. Future improvements include class-weighted training, oversampling, and nonlinear models such as 1D convolutional neural networks.

4. Comparison with original paper

The Logistic Regression classifier serves as the baseline model in this project. It achieved high overall accuracy (approximately 91.5%) on both validation and test sets. However, a detailed analysis of the confusion matrices reveals that this performance is heavily driven by correct classification of the majority class (Class 0).

Minority heartbeat classes, particularly Classes 2 and 3, exhibit substantially lower recall. A large proportion of these samples are misclassified as Class 0, indicating that the linear decision boundaries learned by Logistic Regression fail to separate subtle ECG morphologies associated with rare arrhythmias.

This behavior is expected due to two main factors. First, Logistic Regression assumes linear separability in the standardized feature space, while ECG signals exhibit nonlinear temporal dynamics. Second, the absence of class weighting causes the optimization objective to be dominated by majority-class samples, biasing predictions toward the most frequent category.

Although Logistic Regression provides computational efficiency and interpretability, its limited representational capacity results in poor minority-class generalization, making it insufficient for robust clinical-level ECG classification.