First Practical Machine Learning In Medicine

Hung Viet

***Abstract*—Electrocardiogram (ECG) signal analysis is essential for diagnosing heart diseases. This study applies Machine Learning techniques to classify ECG heartbeat signals into different categories in order to assist medical diagnosis. The ECG Heartbeat Categorization Dataset is used, and data normalization is applied during preprocessing. A K-Nearest Neighbors (KNN) classifier is employed to identify heartbeat patterns based on signal similarity. The model performance is evaluated using standard classification metrics. The results show that the proposed approach can effectively classify heartbeat types and support clinical decision-making.**

I. Introduction

Nowadays, the application of Machine Learning in medicine is becoming increasingly important. In particular, Machine Learning techniques are widely used to analyze heart activity data in order to assist doctors in the diagnosis and treatment process. By learning patterns from medical data, these models can help predict abnormal heart activity and support clinical decision-making.

II. Dataset

The dataset used in this study is the ECG Heartbeat Categorization Dataset, which is publicly available on Kaggle. This dataset is derived from the MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database which are widely used for heartbeat classification research. It contains preprocessed electrocardiogram (ECG) signals representing individual heartbeats.

Each sample in the dataset consists of 187 numerical features, which represent the amplitude values of an ECG waveform over time, and one class label indicating the type of heartbeat. The dataset includes five different heartbeat categories, namely normal beats and four types of abnormal beats. The labels are encoded as integers from 0 to 4.

III. Data Preprocessing

Before training the Machine Learning model, several data preprocessing steps are applied to ensure reliable and accurate classification results. First, the dataset is separated into input features and target labels. The input features consist of 187 numerical values representing ECG signal amplitudes, while the target variable indicates the heartbeat class.

Since the K-Nearest Neighbors (KNN) algorithm is a distance-based method, feature scaling is a critical preprocessing step. In this study, StandardScaler is used to normalize the input features. This method transforms the data so that each feature has a mean value of zero and a standard deviation of one. As a result, all features contribute equally to the distance calculations used by the KNN model.

IV. Exploratory Data Analysis

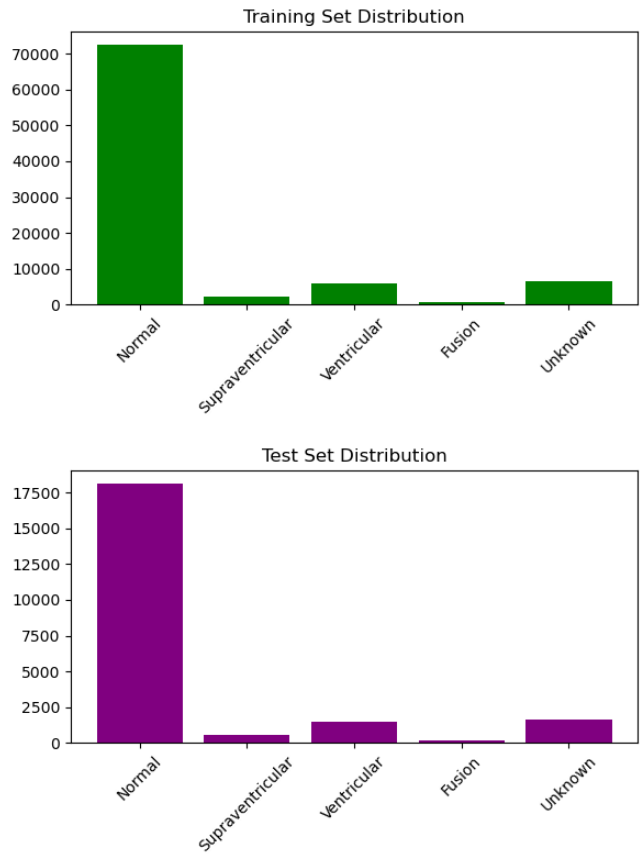

Fig. 1. Class distribution of ECG heartbeat categories in the training set (left) and test set (right).

From the class distribution plots, there is an imbalance between five classes in the dataset. The Normal heartbeat class dominates both the training and test sets, accounting for the majority of samples. In contrast, abnormal heartbeat classes such as Supraventricular, Ventricular, Fusion, and Unknown have significantly fewer samples, with the Fusion class being the smallest.

V. Model Training and Model Evaluation

The KNN model is selected as the primary classification model because KNN is a non-parametric and instance-based learning method that classifies a sample based on the majority class among its nearest neighbors in the feature space.

After data preprocessing and normalization using StandardScaler, the training dataset is used to fit the KNN model. The Euclidean distance metric is employed to measure similarity

between ECG signals, as it effectively captures differences between waveform patterns. The number of neighbors is chosen empirically(number of neighbors = 5), and distance-based weighting is applied to give more influence to closer neighbors during classification.

TABLE I
CLASSIFICATION REPORT OF THE KNN MODEL ON THE TEST DATASET

Class	Precision	Recall	F1-score	Support
Normal	0.98	0.99	0.99	18117
Supraventricular	0.88	0.65	0.75	556
Ventricular	0.93	0.90	0.92	1448
Fusion	0.76	0.65	0.70	162
Unknown	1.00	0.96	0.98	1608
Accuracy		0.97		21891
Macro Avg	0.91	0.83	0.87	21891
Weighted Avg	0.97	0.97	0.97	21891

The model achieves an overall accuracy of 97.4%, indicating strong classification capability. The results show excellent performance in identifying the Normal heartbeat class, which has the highest number of samples. High precision and recall values are also observed for the Ventricular and Unknown heartbeat classes.

However, lower recall values are observed for minority classes such as Supraventricular and Fusion heartbeats. This outcome is mainly due to the class imbalance in the dataset, which makes minority class detection more challenging. To address this issue, macro-averaged metrics are also reported to provide a fair evaluation across all classes.

Overall, the experimental results demonstrate that the KNN model performs effectively for ECG heartbeat classification and has potential to support medical diagnosis, while also highlighting the importance of handling class imbalance in future improvements.