Classification ECG Signals Base on k-Nearest Neighbors (k-NN) Algorithm

Wameedh Raad Fathel [1]
[1]Ministry of Education/ General Directorate of Education in Nineveh/Iraq
Wamed81@gmail.com

Marwa Mawfaq Mohamedsheet Al-Hatab [2]
[2 3]Technical Engineering College /Northren Technical University / Mosul / Iraq
marwa.alhatab@ntu.edu.iq

Maysaloon Abed Qasim [3]
[2 3]Technical Engineering College /Northren Technical University / Mosul / Iraq
maysloon.alhashim@ntu.edu.iq

ABSTRACT

Abnormal cardiac rhythm known as atrial fibrillation (AF) is marked by an atria's fast and erratic pulse. It often begins in short periods of abnormal beating which becomes longer and may be constant over time. Usually it presents no symptoms and a typical ECG affected by Atrial Fibrillation does not present any P wave and shows an irregular ventricular rate. In this study, the k-Nearest Neighbors (K-NN) algorithm has been used to classifier 5000 samples of cardiac signals. After preprocessing the data, it was split into the three classes represented, namely: Normal (N), AF, and Noisy Rhythm (NR). In a ratio of 1:1, the data were split into two groups: training dataset and test dataset, to perform the classification. It was obtained from the dataset, the highest sensitivity recorded for N cases is **92**% and the highest specificity recorded for AF is 99%. The classification accuracy obtained is 90% and the value for area under the curve (AUC) is 0.94.

Keywords:

1. Introduction

The heart beats in a steady sinus rhythm between 60 and 100 beats per minute; the trace represents a single pulse and lasts around 0.8 seconds. Each trace includes events that are 'P', 'Q', 'R', 'S', and T wave-classified by convention. The 'P' wave's accompanying atrial depolarization and contraction [1]

The "T" wave is a visual representation of the ventricles' repolarization and relaxation. The sinus node, which functions as the base maker and typically initiates the action potential in the heart, regulates the contraction of cardiac muscle [2]. Regular Heart Beats (HBs) occur while the heart contracts and relaxes. Atrial fibrillation (AF) prevents the heart's upper chambers, the atria, from properly pumping blood into the ventricles. To define AF as an unstable or quivering rhythm (arrhythmia) that may result in stroke, blood clots, heart failure and other heart-related conditions. Rapidly firing action potentials in the atrium or pulmonary veins that are chaotically firing cause it to occur. An atrial rate between 400 and 600 beats per minute (bpm) is the result [3]. Due to the high atrial rate and limited amplitude of the action potentials produced, 'P' waves are not visible on the electrocardiogram (ECG) in people with AF [4]. Despite the potential of noise in the ECG trace, the central problem with machine classification is the significant diversity in the shape of beats pertaining to the same class and beats of similar form belonging to different classes. Auto-detecting the changes in the ECG signal due to AF represents a cardiology challenge [5]. As a consequence, the three processes that make up computer-based diagnosis approaches are as follows: EGC beat recognition, beat feature extraction, and beat classification. The feature extraction can be carried out statistically, in the frequency domain, or in the temporal domain. Because they are adaptable and extremely effective, statistical characteristics in the temporal domain were chosen. The decomposition is generated from the signal itself since it is based on the local characteristic time scale of the data. The number of ‘P' waves will decrease, as was previously seen in AF, and the space between consecutive R waves will rise [6]. Based on these facts, a function to calculate the inter-beat interval was developed and put to use together with other characteristics like counting the amount of 'P' waves, which should serve as a reliable indicator of cases of AF, and monitoring variations in heart rate. The purpose of this work is to offer a useful technique for categorizing patients into Normal (N), Noise Rhythm (NR), and AF groups using the k-nearest neighbor’s algorithm (k-NN).

2. Methods

2.1. Datasets

The program used information from the PhysioNet 2017 Challenge Database [7], which gathered 5000 signals and distributed them as follows: 12% AF, 83% N, and 5% NR. In order to complete all the classification phases, three matrices of each data type were produced and divided into training and testing data in the ratio of 1:1.

2.2 Feature Extraction

From the signal amplitude, we derive several power spectrum properties that served as features of the ECG signal.

- The typical signal's spectral power
- Variation in signal spectral power
- Determined the root mean square fluctuation of the integrated and detruded time series. The raw ECG time series may now be seen to include long-range correlations. The raw signal is used to calculate the quick Fourier transform.
- Time series' root mean square variation
- Time series' average total power
- Variance in the temporal series' overall power

2.3 K-Nearest Neighbors Classifier

Recently, artificial intelligence and modern technologies included in health applications [8-11]. An easy-to-use supervised Machine Learning (ML) approach that may be used to tackle both classification and regression issue [12]. KNN is the algorithm used by the majority of ML algorithms [13]. The most frequent class label point in the space is allocated to a particular class k closest training sets classifier. In the KNN classifier the Euclidean distance is computed between two points (samples). Euclidean distance between the two positions $|\rho^d - \rho^g|$ in dimensional z and $z+1$ might be the embedding dimension l and delay time T_l these distances are then provided by [14].

$$\epsilon^2_z = \sum_{l=0}^{z-1} [\rho^d(t+lT_l) - \rho^g(t+lT_l)]^2 \quad (1)$$

Obtained features are created throughout the classification training process and expressed in the allocated sample. K closest samples are chosen after distances between the new vector and all combined vectors are calculated.

2.4 Statistical analysis

In order to validate the classifier for each classification process, the positive rate (Sensitivity) and the false positive rate (Specificity) are computed [15, 16, 17]. Furthermore, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) depict the connection between sensitivity and specificity for different split points [18, 19]. Sensitivity is the proportion of individuals that tests positive and exhibits the desired condition. Specificity is the proportion of individuals who do not has the specified disease and has negative test results [20, 21]. The model's ability to discriminate between groups is measured by the AUC and the greater the AUC, the better the ability of the model to separate people with and without illness. The sensitivity/specificity connection associated to a decision threshold is shown by each point on the ROC curve.

3. Results

In this study, classification was performed using dataset consists of 5000 samples for three groups of N, AF and NR. Various parameters were measured to validate the classification procedure. The data division has been approved into two parts by 50% for training and 50% for testing. Dell, Corei5 and RAM -16 computer type has been used during training and testing. According to Confusion matrixes in figure (1) the true and predicted class for each CLASS 0_N, CLASS 1__AF and CLASS 2_ NR can be observed.

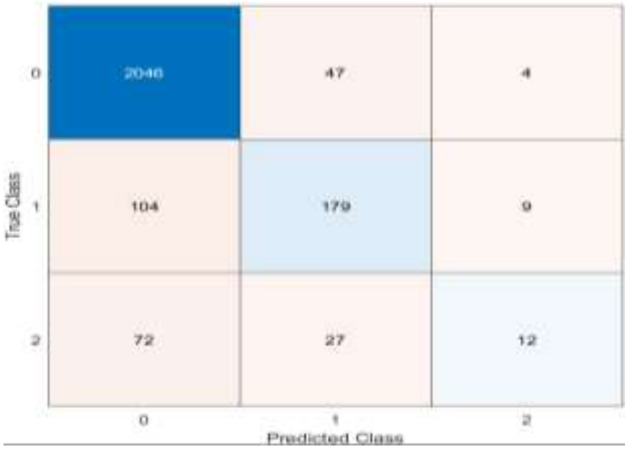

Figure 1: Confusion matrix of all cases (N, AF AND NR) for dataset

By summing and combining Cell (C) values, we can calculate N, AF and NR classes, table 1 shows True Negative (TN), False Positive (FP), True Positive (TP), and False Negative (FN) values.

Table1: The N, AF, and NR cell values for ECG signals are summed.

CLASS 0_N	CLASS 1__AF	CLASS 2_ NR
TP=C1	TP=C5	TP=C9
FP=C2+C3	FP=C4+C6	FP=C7+C8
TN=C5+C6+C8+C9	TN=C1+C3+C7+C9	TN=C1+C2+C4+C5
FN=C4+C7	FN=C2+C8	FN=C3+C6

The following formula has been adopted and applied to the three categories individually used equation 3 for sensitivity calculation and equation 4 for specificity calculation base on Table (1). Table (2) shows the sensitivity and specificity

$$\text{Sensitivity} = TP / (TP + FN) \qquad (2)$$

$$\text{Specificity} = TN / (TN + FP) = 1 - \text{False positive rate} \qquad (3)$$

Table 2: Sensitivity and specificity calculation for N, AF, and NR cases

	CLASS 0_N	CLASS 1__AF	CLASS 2_NR
Sensitivity	92%	71%	48%
Specificity	82%	99%	96%

In order for a test to appropriately categorize a person as having an AF, N, or NR sensitivity and specificity were initially considered in this study. The sensitivity of a test refers to its capacity to provide a positive finding for someone who has a condition. Few false negative findings and fewer missing N cases has been obtained. so 92% highly sensitive for CLASS 0_N case was recorded. While the system recorded lower sensitivity for detecting cases of CLASS 1_ AF and CLASS 2_ NR, the sensitivity for CLASS 1_ AF and CLASS 2_ NR were 71% and 48% respectively. A test's specificity is its capacity to label someone as negative for an illness if they do not have it. The top outcomes from our methodology were 82% for CLASS 0_ N, 99% for CLASS 1_ AF, and 96% for CLASS 2_ NR.

secondly, the performance of model can be evaluated base on parameters shown in figure 2. Using the four variables True Positive Rate (TPR), Positive Predictive Values (PPV), False Negative Rates (FNR), and False Discovery Rates (FDR), (TP, FP, TN, and FN) values can be evaluated.

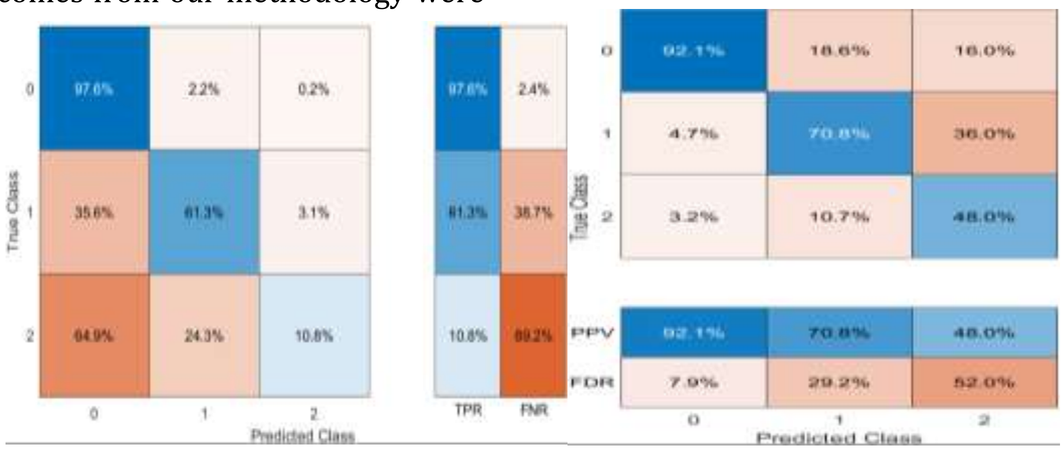

Figure 2: illustrate the TPR,FNR, PPV and FDR values for N AF and NR for ECG signals

According to figure, the proposed system proved the good performance even if the data is unbalanced, because in most of the cases the TPR and PPV values were high compared to the FNR and FDR values which were low. 97%, 61%, and 10% successive values of N, AF, and NR for the TPR condition have been obtained. While rates of 92.1%, 70% and 48% for the three cases, respectively have been recorded for PPV condition

Theridly AUR is another measure of test performance ,The relationship between sensitivity and specificity for set of tests. ROC curves are frequently used to graphically represent this relationship The benefits of using this test are also revealed by the ROC curve's area under the curve. In the system , figure 3 shows the the discrimintation of the sysyte is very high because the AUC value is 0.94 which represents the system ability to diagnose patients with all cases and according to this value the result is excellent.

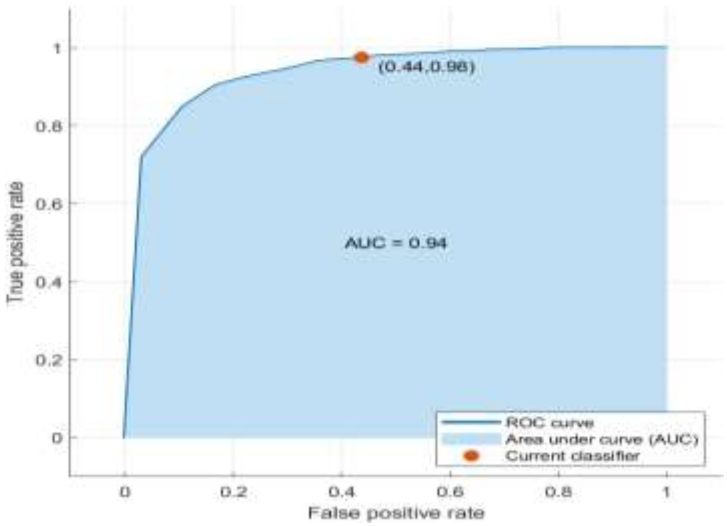

Figure 3: The ROC curve and AUC for classification N, AF and NR system

4. Conclusion

Generally, it is recommended to have a highly sensitive, high specificity test. This is often not possible. Usually there is a trade-off. In this work, by relying on 5,000 clinical test samples, system has been proven the efficiency of its classification of the three heart signals using KNN and reached 90% and 0.94 value for AUC. With an examination of sensitivity and specificity, which gave excellent results with the classification of the N signal, as it where record 92%. While the AF signal was distinguished, the specificity reached 99%. Finally; The system was enhanced by recognizing NR signals with a specificity of 96%.

Reference

1. Fitzgerald, Bishop Ernest A. "Basic Cardiac Monitoring: The Electrocardiogram." *Introduction to Radiologic and Imaging Sciences and Patient Care-E-Book* (2015): 200.
2. Zhang, H., et al. "Mathematical models of action potentials in the periphery and center of the rabbit sinoatrial node." *American Journal of Physiology-Heart and Circulatory Physiology* 279.1 (2000): H397-H421.
3. Sun, Hui, et al. "Cellular mechanisms of atrial contractile dysfunction caused by sustained atrial tachycardia." *Circulation* 98.7 (1998): 719-727.
4. Li, Zicong, and Henggui Zhang. "Automatic Detection for Multi-Labeled Cardiac Arrhythmia Based on Frame Blocking Preprocessing and Residual Networks." *Frontiers in cardiovascular medicine* 8 (2021): 616585.
5. Jiang, Kai, et al. "High accuracy in automatic detection of atrial fibrillation for Holter monitoring." *Journal of Zhejiang University SCIENCE B* 13.9 (2012): 751-756.
6. Abdukalikova, Anara. "Machine Learning assisted system for the resource-constrained atrial fibrillation detection from short single-lead ECG signals." (2018).
7. Clifford GD, Liu C, Moody B, Silva I, Li Q, Johnson AEW, Mark RG. AF Classification from a Short Single Lead ECG Recording: the PhysioNet Computing in Cardiology Challenge 2017. In Computing in Cardiology. Rennes (France), 2017;
8. AL-Hatab, M. M., et al. "Comparison study between three axis views of vision, motor and pre-frontal brain activities." *Journal of Critical Reviews* 7.5 (2020): 2598-2607.
9. AL-Hatab, M. M., et al. "Classifying various brain activities by exploiting deep learning techniques and genetic algorithm fusion method." *TEST Engineering & Management* 83 (2020): 3035-3052.
10. Al-Hatab, Marwa Mawfaq Mohamedsheet, Raid Rafi Omar Al-Nima, and Maysaloon Abed Qasim. "Classifying healthy and infected Covid-19 cases by employing CT scan images." *Bulletin of Electrical Engineering and Informatics* 11.6 (2022): 3279-3287.
11. Al-Hatab, Marwa Mawfaq Mohamedsheet, et al. "Innovative Non-Invasive Blood Sugar Level Monitoring for Diabetes Using UWB Sensor." *Journal of Optoelectronics Laser* 41.4 (2022): 422-437.
12. Uddin, Shahadat, et al. "Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction." *Scientific Reports* 12.1 (2022): 1-11.
13. Du, Yuqing, et al. "Automatic recognition of space-time constellations by learning on the Grassmann manifold." *IEEE Transactions on Signal Processing* 66.22 (2018): 6031-6046.
14. Lalkhen, Abdul Ghaaliq, and Anthony McCluskey. "Clinical tests: sensitivity and specificity." *Continuing education in*

anaesthesia critical care & pain 8.6 (2008): 221-223.

15. Head, J. F., C. A. Lipari, and R. L. Elliott. "Comparison of mammography and breast infrared imaging: sensitivity, specificity, false negatives, false positives, positive predictive value and negative predictive value." *Proceedings of the First Joint BMES/EMBS Conference. 1999 IEEE Engineering in Medicine and Biology 21st Annual Conference and the 1999 Annual Fall Meeting of the Biomedical Engineering Society (Cat. N*. Vol. 2. IEEE, 1999.
16. Shreffler, Jacob, and Martin R. Huecker. "Diagnostic testing accuracy: Sensitivity, specificity, predictive values and likelihood ratios." *StatPearls [Internet]*. StatPearls Publishing, 2022.
17. Nahm, Francis Sahngun. "Receiver operating characteristic curve: overview and practical use for clinicians." *Korean journal of anesthesiology* 75.1 (2022): 25-36.
18. Ying, Gui-Shuang, et al. "Tutorial on Biostatistics: Receiver-Operating Characteristic (ROC) Analysis for Correlated Eye Data." *Ophthalmic Epidemiology* 29.2 (2022): 117-127.
19. Meyer, Ingmar Sören, and Florian Leuschner. "The role of Wnt signaling in the healing myocardium: a focus on cell specificity." *Basic Research in Cardiology* 113.6 (2018): 1-12..
20. Grunau, Gilat, and Shai Linn. "Commentary: sensitivity, specificity, and predictive values: foundations, pliabilities, and pitfalls in research and practice." *Frontiers in Public Health* 6 (2018): 256.