EARLY PREDICTION OF HEART DISEASES USING MACHINE LEARNING

Dilip Uike

PG Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India

K. P. Wagh

PG Department of Computer Science, Sant Gadge Baba Amravati University, Amravati, Maharashtra, India

Abstract: *Now a day's heart disease becomes a common in any age group. Severe heart attack occurring in very young and highly performing athlete. To predict heart related problem in real time are very much challenging task. If it is able to predict heart related ailment in real time frame then it may help to save someone's life. It is mandatory to develop a technique to accurately predict the symptoms and classify that in particular heart disease problem, so as to take quick action to proceed for treatment. Here aim is to develop a system with low computational cost and best performance. It is difficult to identify heart disease because of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, abnormal pulse rate and many other factors. Various techniques in data mining and neural networks have been developed to find out the severity of heart disease among humans. The nature of heart disease is complex and hence, the disease must be handled carefully, if not doing so may affect the heart or cause premature death. The perspective of medical science and data mining with machine learning are used for discovering various sorts of metabolic syndromes. Data mining with classification plays a significant role in the prediction of heart disease and data investigation. Here the features are extracted and classified using ANN and Machine learning methods. It is important to save someone's life by predicting heart problem early & accurately.*

Keywords: **Data Mining, Heart Disease, Machine Learning, Neural Network, Prediction**

1. INTRODUCTION

Arrhythmia Classification plays a major role while diagnosing heart diseases. Any change in the regular sequence of electric impulses is called as arrhythmia. Identifying arrhythmia as early as possible helps the patient in choosing appropriate treatment. Classification of ECG arrhythmia with high accuracy is a challenging problem. Arrhythmia classification requires preprocessing of ECG Signal, extraction of features, and optimization of the features and classification of arrhythmia. As we know Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning (ML) has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. We have also seen ML techniques being used in recent developments in different areas of the Internet of Things (IoT). Various studies give only a basic into predicting heart disease with ML techniques. In this research it has been proposed a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques.

In this method enhanced performance level with high accuracy level of more than 90% through the prediction model for heart disease with the hybrid according to the World Health Organization (WHO), 31% of all global deaths are due to the Cardiovascular Diseases (CVDs). Diagnosing CVDs and ensuring the patients can receive appropriate treatment are necessary to prevent premature deaths. The analysis of the ECG is widely applied in the diagnosis of these heart disorders. The most useful information in the ECG is normally derived from the amplitudes and intervals of these dividual waves that are defined by the fiducial points (e.g. onset, offset, peak). In general, these features are used

to classify the normal and abnormal heartbeats in this process of diagnosis of a specific heart disease, e.g. congestive heart failure (CHF) and cardiac arrhythmia. Therefore, it is necessary to extract various features of ECG in order to diagnose the heart diseases. Among the ECG wave, the QRS complex is relatively easy to identify because of its specific morphology and high amplitude. However, the T-wave delineation is a more challenging task, due to its low amplitude and possibly irregular morphology. In addition, noises such as baseline wandering and power line interference are main factors that can result in faulty T-wave delineation. Over the decade, a number of automated algorithms have been developed for ECG delineation. In general, there are two main groups of ECG feature extraction algorithms, which are QRS detection and non-QRS delineation algorithms. The first QRS detection algorithm was introduced by Pan and Tompkins. There are other attempts for QRS detection based on Shannon energy envelope (SEE), wavelet transform (WT), phase-space reconstruction (PSR), Optimized adaptive thresholding, iterative state machines, and moving-average filters. Concerning the non-QRS delineation algorithms, the main objective is to determine the peaks and boundaries of the individual QRS complexes, P and T waves. The existing literature on ECG wave delineation algorithms is extensive and focuses particularly on frequency aspect, e.g. DWT, the combination of wavelet transform and hybrid hidden Markov models. Among other popular methods, the phasor transform, moving average filters morphological mathematical filtering with Elgendi's algorithm and the correlation analysis based method have also been applied to detect ECG fiducial points. However, the major concern associated with these algorithms is their detection accuracy, more importantly low positive predictivity (+P) caused by the large number of false-positives (FPs) of R-peak detection. In addition, current methods can be error-prone, especially cannot achieve satisfactory performance for T-wave detection due to the variable morphology of the T-wave.

For arrhythmia detection and classification, a number of methods have already been presented in the literature ranging from the traditional feature-based machine learning process to the end-to-end deep learning process in recent times. In feature-based arrhythmia detection techniques, various feature extraction approaches are employed, such as wavelet transform, principal component analysis, independent component analysis and Hermite function. For performing classification with the extracted features, support vector machine (SVM), K-nearest neighbour, feed-forward neural network and random forest have been used. These approaches mostly depend on handcrafted feature extraction process that most often leads to loss of information required for the classification due to improperly chosen features or inadequate features. Automating the process of feature extraction and classification was the primary motivation behind the popularity of end-to-end deep learning-based frameworks.

2. Related Work

Up till now what work has done to related work of various methods have been used for knowledge discovery by using known methods of data mining for prediction of heart disease. In this work, more innovation has been carried out to produce a prediction model using not only distinct techniques but also by relating & combining two or more classification algorithms. Methods used to analyze the performance are also discussed. Classification of electrocardiogram (ECG) signals is obligatory for the automatic diagnosis of cardiovascular disease. With the recent advancement of low-cost wearable ECG device, it becomes more feasible to utilize ECG for cardiac arrhythmia classification in daily life. In this research it has been proposed an approach to classify five types of cardiac arrhythmia, namely, normal beat (N), atrial premature contraction (A), premature ventricular contraction (V), left bundle branch block beat (L), and right bundle branch block beat (R). The combined method of frequency analysis and Shannon entropy is applied to extract

appropriate statistical features. Information gain criterion is employed to select features that the results show that 10 highly effective features can obtain performance measures comparable to those obtained by using the complete features. The selected features are then fed to the input of Random Forest, K-Nearest Neighbor, and J48 for classification. To evaluate classification performance, tenfold cross validation is used to verify the effectiveness of our method. Experimental results show that Random Forest classifier demonstrates significant performance with very high accuracy. The major challenge that the Healthcare industry faces now-a-days is superiority of facility. Diagnosing the disease correctly & providing effective treatment to patients will define the quality of service. Poor diagnosis causes disastrous consequences that are not accepted. Records or data of medical history is very large, but these are from many dissimilar foundations. The interpretations that are done by physicians are essential components of these data. The data in real world might be noisy, incomplete and inconsistent, so data preprocessing will be required in directive to fill the omitted values in the database. Even if cardiovascular diseases is found as the important source of death in world in ancient years, these have been announced as the most avoidable and manageable diseases. The whole and accurate management of a disease rest on the well-timed judgment of that disease. A correct and methodical tool for recognizing high-risk patients and revised data mining data for timely analysis of heart infection looks a serious problem. Different person body can show different symptoms of heart disease which may vary accordingly. Though, they frequently include back pain, jaw pain, neck pain, stomach disorders, and tininess of breath, chest pain, arms and shoulders pains. There are a variety of different heart diseases which includes heart failure and stroke and coronary artery disease. Even though heart disease is acknowledged as the supreme chronic sort of disease in the world.

A healthy way of life and timely analysis are the two major factors can suppress origins of heart disease. Conducting steady check-ups shows outstanding role in the judgment and early prevention of heart disease difficulties. Several tests comprising of angiography, chest X-rays, echocardiography and exercise tolerance test support to this significant issue. These tests are expensive and involve availability of accurate medical equipment. As the method discussing here is used to predict the particular disease related to heart. Heart failure is one of the serious problem that we have to deal with, because it exist sometimes unexpectedly without experienced by patient and may leads to death within short time. Here we are building system in such a way that it would give a perfect prediction about particular heart related ailment going to happen, so as to take remedies within time limit and save someone's life. On the basis of its high performance and accuracy it is considered as reliable product that may help people to diagnose earlier.

There are many types of heart diseases with different risk of handling. If we predict perfectly and within time limit it may help to save the expenses that we are going to spend on patient. This research deals with hybrid technology that combines two or more than two algorithms to increase accuracy and performance of system. In future if required given product can be converted to another form by making some changes or again combining different algorithms to make it feasible product.

3. DETECTION AND CLASSIFICATION OF ECG

The overall methodology of the proposed system is shown in Figure 1, which consists of (1) pre-processing implemented by filtering and segmentation, (2) one-dimensional ECG signal to spectro-temporal images conversion achieved by wavelet and short-time Fourier transform, (3) feature auto extraction and classification fulfilled by multi-channel dense convolutional neural networks. The following contents introduce those components section by section.

First of all, the target ECG signals are resampled to 200 Hz as input for the proposed system. Next, bandpass filters, notched filters and adaptive filters are applied to reconstruct clean signals and remove noises from the ECG signal, including the power-line interference, baseline wander, muscle contraction noise, etc. Then, the proposed system segments the ECG signal into slices in 1s, 3s, and 6s based on the locations of detected R-peaks

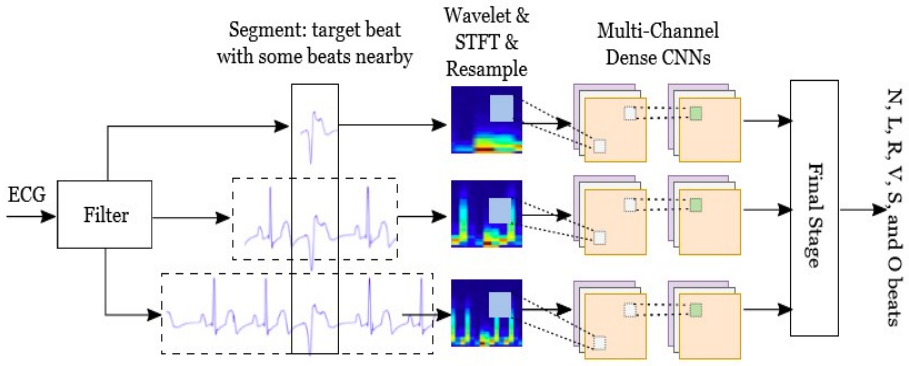

Figure 1. The overall methodology of the proposed system for the proposed ECG beat classification. The system consists of pre-processing (filtering and segmentation), feature extraction (wavelet and short-time Fourier transform) and classification (dense CNNs). It has three channels of dense CNNs where each channel detects the same target beat but using different windows to obtain various beat-to-beat information represented by wavelets transform and short-time Fourier transform.

Wavelet transform is widely used to describe the ECG morphology features since it's robust to noise and can effectively represent both time and frequency information in different resolutions. In practice, the digital wavelets transform with a specific wavelet family is always selected in applications. In our proposed system, we select Daubechies 4 as the mother wavelet and choose 4th level WT components as the coefficients to describe the abnormalities in ECG beats as recommended in. In this way, some particular beats which include wide QRS morphologies or more low-frequency spectrums (e.g. bundle branch block beat and ventricular ectopic beat) will be significantly amplified or enlarged by WT while normal beats keep the same. This characteristics of WT distinguishes some special beats and make them easy to-detect.

Although pre-processed ECG segments can be directly used as input of one-dimensional CNN to predict types of heartbeats, time-frequency representation may be a better choice inspired by speech recognition applications where spectrogram features of speech are applied to improve the performance on traditional CNN. This proposed system converts the wavelet transformed ECG segments into time-frequency domain by using the short-time Fourier transform and obtain the ECG spectrogram representation.

Mathematically, it can be described as follows:

$$s_i(k,m) = \sum_{n=0}^{N-1} x_i(n) w(m-n) e^{-j\frac{2\pi}{N}kn},$$

Where w(•) is the window function, e.g., Hamming window. Si (k,m) is the spectrogram of xi, which has a two-dimension structure. In the proposed system, temporal-spectro images are first generated by the WT and STFT from 1s, 3s, and 6s ECG segments, then sent to multichannel dense CNNs for further analysis.

For ECG signals, directly applying one-dimensional deep neural networks is promising but lacking noise-robust features, while existing CNNs based ECG beat classifiers reaches their bottlenecks since they fail to consider beat-to-beat information together with the target single-beat morphologies. To solve this issue proposed approach uses multiple

CNNs to receive all the spectro-temporal images extracted from ECG segments and combine that information for beat detection. Specifically, each CNN analyzes one group of spectro temporal images converted from different time-scaled ECG segments (1s, 3s or 6s). Its output is finally merged to the final stage as the prediction of the heartbeats. The combination in the final stage merges the predictive information of single-beat morphologies with beat-to-beat information, which increases the performance on ECG beat classification. To implement those CNNs, this paper deploys multiple dense connected convolutional networks (dense CNN) with the same configuration as shown in Figure 2. The dense CNN is famous for its outstanding ability to solve the vanishing gradient issue and now widely applied in image processing. Each dense CNN can be seen as refined.

Multiple dense CNNs are finally linked at the final stage, which combines and processes CNN's outputs with maximum pooling, average pooling, and softmax function. In this way, the combination of all dense CNNs' outputs can cover spectro-temporal information not only from single-beat morphologies but also beat-to-beat characteristics, resulting in an enhanced beat-by-beat classification performance.

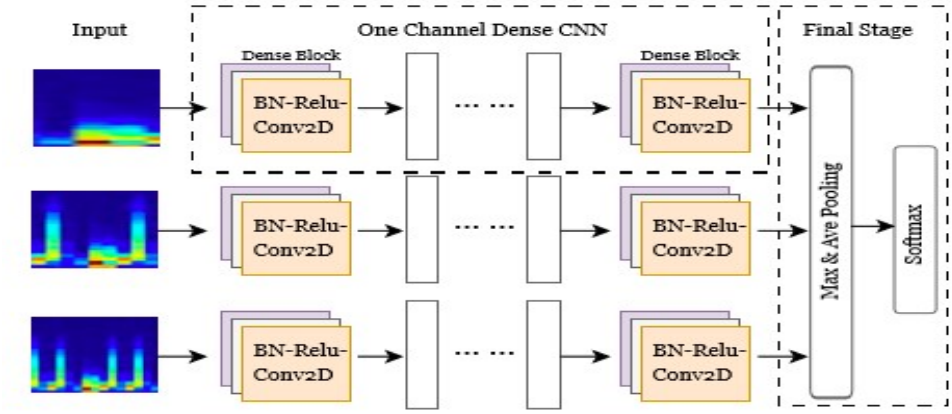

Figure 2. The detailed structures of multi-channel dense CNNs and the final stage. For each channel of a dense CNN, every two and preceding layers are explicitly connected in a dense block, and the outputs of all dense CNNs are combined and processed at the final stage

4. Detection of P, R, and T Waves Peak

In this section, we will present the proposed method based on the combination of hierarchical clustering and DWT. Our algorithm aims at detection of QRS complexes and T peaks from the sequence of successive ECG signals. In this work, in order to analyze at least one R-peak and at most two R- peaks each step, we set a sliding time window of 1.2s at each step, as we experimentally found out that this is the optimal time window for detecting the R-peaks. Our proposed technique is structured as a four-stage process. First is ECG pre-processing. Raw ECG signals were filtered using fourth- order Butterworth high-pass filter and low-pass filter with the corresponding cut-off frequency of 1 Hz and 30 Hz to remove the noise and baseline wandering. After that, the ECG signal was normalized so that all the values will be in the range [0, 1]. An example can be seen in Figure 3.

Second stage is using hierarchical clustering to determine the R-clusters and non-R clusters, then, identify the R-peaks from R-clusters. The third stage pertains to the T-wave boundary detection based on the R-peak and an ECG period template. The final stage is to find the T-peaks by using DWT and MMA.

$$\tilde{E}(n) = (E(n) - E_{min}) / (E_{max} - E_{min})$$

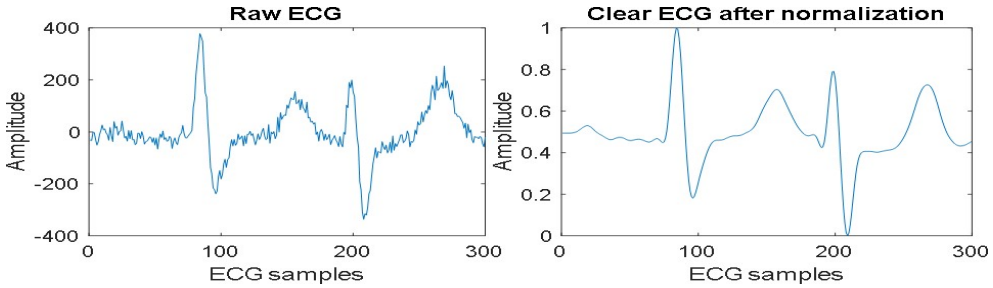

Figure 3. An example of ECG pre-processing (annotation: se1891m from QT database. The sampling frequency is 250Hz hence a time window has 300 ECG samples)

The strategy of the proposed algorithm is to first identify the R-peak position by using hierarchical clustering. In consider every ECG sample as an individual cluster initially and all R wave samples are merged into a cluster based on their similarity. We select Euclidean distance to measure the similarity between each ECG sample. Next, we need to select a clustering method to determine the R-wave clusters.

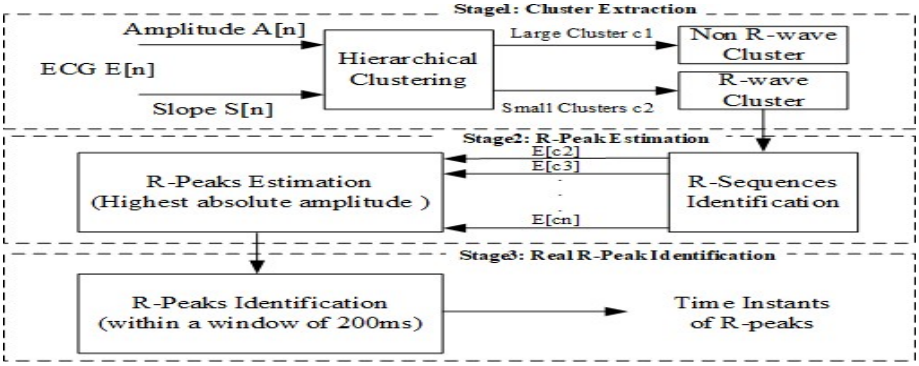

Figure 4. An example of ECG R-wave cluster extraction

In this case, single and complete linkage algorithms are not Applicable because they reduce the assessment of the cluster quality to a single similarity between a pair of objects and they cannot fully reflect the distribution of objects in one cluster. Therefore, they usually produce an undesirable cluster that may include non-R samples. The average linkage algorithm can avoid this situation because it determines the clusters based on the average distance between all pairs of objects. The process of our R-peak detector can be divided into three main stages, namely cluster extraction, R-peak estimation and real R-peak identification. The block diagram of the proposed R-peak detector is drawn in Figure 4. First, for ECG data E[n], each ECG sample E[i] is considered as an individual object with its amplitude A[i] and slope S[i]. The slope of an object is defined as the average absolute value of the amplitude difference between E[i] and E[i + 1]; and E[i 1] respectively. The expression is shown in (1). Then, these objects are considered as the input data to a two- dimensional hierarchical clustering system. Then, the hierarchical clustering system will calculate the distance between each object (ECG sample) using (1). Initially, two ECG samples with the shortest distance are merged into a cluster. Then the hierarchical clustering system selects two clusters with the shortest average distance of the ECG samples between them by (6), groups them together into a new cluster and repeats the procedure with the remaining ECG samples. Until the number of cluster become 2, one is a large cluster and another is a small cluster. The cluster with small number of object is R-cluster, since the ECG samples of R-wave is just a small part of the ECG signal. Other ECG samples are non-R cluster. Here, some of the ECG samples with

large amplitude are still considered as non-R cluster, that is because the slope of these objects are relatively low. This is particularly important to solve T-wave oversensing.

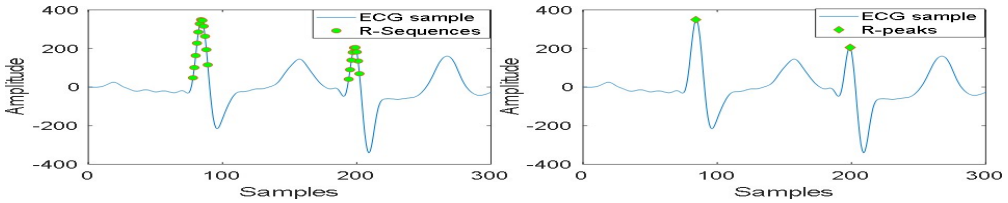

Figure 5. An example of R-peak Estimation

Once the R-wave clusters are determined in the last stage, the next step is to identify the R-wave ECG sample sequences from these R-wave clusters. There are two R-wave sequences (Green points) from the R- peak cluster which are identified using hierarchical clustering. Then, this step is to find the ECG sample with highest absolute value of the amplitude within each R-wave sequences and Consider these ECG samples are R-peaks. The final result is shown on the right of the Figure 5.

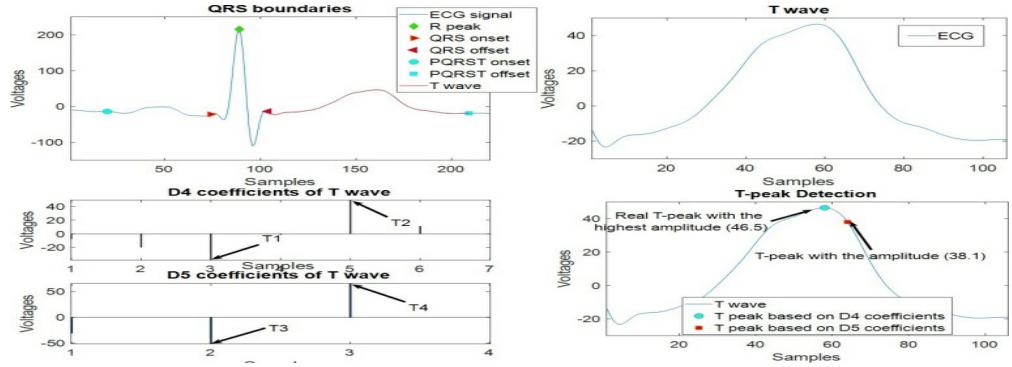

Figure 6. T peak detection based on DWT and MMA

The green diamond is marked as R-peak (t0). Moreover, the PQRST-onset (offset) and QRS-onset (offset) are t1 (t2) and t3 (t4) respectively. In this case, [t4, t2] is considered as T wave boundary of the template. Then we calculated the distance between R-peak and QRS offset, PQRST offset as t4 t0 and t2 t0 respectively. Once we obtained the R-peak (tR), the QRS offset (T onset) and PQRST offset (T offset) is initially estimated as tT on = tR + (t4 t0) and tT off = tR + (t2 t0) respectively. The next step is to calculate the Mean Square Error (MSE) between ECG samples (Ei) in the T wave boundary of the template and ECG samples (E˜i) in each estimated T wave boundary using (10), where n is the number of ECG samples in T wave boundary. Finally each estimated T wave boundary is moved forward and backward by 20 samples E˜i±20 and the respective MSE is calculated. The final T wave boundaries are extracted based on the minimum MSE between Ei and E˜i±20. It also need to be noted that our obtained T wave boundaries obtained here may not be the actual T wave boundaries, since once the width of ECG beat become abnormal, such T wave boundaries cannot be identified accurately using a normal template. Therefore, the main idea of our algorithm is to determine the approximate range based on the T boundary to make sure the T-peak is in this range. It can be seen in Figure 6, the red points are QRS offset (T onset) and yellow points are PQRST offset (T offset).

5. Result and Analysis

The proposed approach is implemented in Python and Pytorch. In the test, ECG data sampled at 360 Hz from the MIT-BIH arrhythmias database (MITdb) is downloaded and then resembled at 200 Hz as input signals. For this target database, the entire 44 patients' ECG records are collected from the lead II for testing the performance of the proposed system against other existing works. In this paper, we

describe the performances using sensitivity (Se) and positive predictive value (PPV) on target heartbeats as defined in Table II. The proposed algorithm shows the high performance in the delineation of ECG R and T-peaks. Concerning the R peak detection, method achieves 99.89% Se, 99.94% +P and 99.83% Acc for R-peak detection. The performance in terms of +P and Acc is higher than other previously proposed works. it can be seen in our work, false positives of R-peak detection can be effectively reduced and only exist in the records with baseline drifts. In the future work, the performance of our algorithm can be improved with the removal of baseline drifts in such ECG records. Furthermore, this algorithm achieves 100% Se and 99.83% +P over the manually annotated QT database. It also achieved a Se of 99.92% and a +P of 99.96% over the automatically annotated QT database. The number of true positive of our algorithm for R- peak detection is larger than these studies. Overall, the R-peak detection performance of our algorithm is generally better or comparable with the previous work. The future work of the additional classes of arrhythmia signals for better analysis. In the future these implementation methods can be adapted in to point of care type devices. Our method may have problems in T peak detection for "sudden death" records because the morphology of T waves in these records may be erratic. Further study is needed to investigate this type of condition.

6. Conclusion

We proposed a novel ECG beat classifier which use the multi-channel neural network and spectro temporal features to identify heartbeat. It first converts one dimensional ECG into spectro-temporal images via wavelets transform and STFT and then applies multiple dense convolutional neural network to further process both beat-to-beat information and single-beat morphologies for automatic heartbeat classification. Based on the comparison with the other five widely used ECG beat classifiers, our proposed system has shown its ability to consistently produce the best overall performance for ECG beat identification. Here we have proposed a new algorithm based on the hierarchical clustering and discrete wavelet transform for the automatic delineation of the ECG fiducial points (R and T peaks). The use of hierarchical clustering allows for identifying the R clusters and determining the R peaks with high accuracy. The combination of DWT and MMA analysis help us to detect the T peaks with high sensitivity. Our algorithm has been validated on the MIT-BIH arrhythmia database and QT database. The results show that our algorithm can solve T-wave oversensing problem and effectively reduce the number of R-peak false-positive detection. Moreover, the performance of algorithm is generally better than other referenced algorithms.

REFERENCES

[1] *D. M. S. Manikandan and K. Soman, "A novel method for detecting rpeaks in electrocardiogram (ecg) signal," Biomedical Signal Processing and Control, vol. 7, no. 2, pp. 118–128, 2012.*

[2] *N. Thiamchoo and P. Phukpattaranont, "R peak detection algorithm based on continuous wavelet transform and shannon energy," ECTI Transactions on Computer and Information Technology (ECTI-CIT),vol. 10, no. 2, pp. 167–175, 2016.*

[3] *M. Elgendi, "Fast qrs detection with an optimized knowledge-based method: Evaluation on 11 standard ecg databases," PloS one, vol. 8, no. 9, p. e73557, 2013.*

[4] *M. Cesari, J. Mehlsen, A.-B. Mehlsen, and H. B. D. Sorensen, "A new wavelet-based ecg delineator for the evaluation of the ventricular innervation," IEEE journal of translational engineering in health andmedicine, vol. 5, pp. 1–15, 2017.*

[5] *A. Mart´ınez, R. Alcaraz, and J. J. Rieta, "A new method for automatic delineation of ecg fiducial points based on the phasor transform," in 2010 Annual International Conference of the IEEE Engineering in Medicineand Biology. IEEE, 2010, pp. 4586–4589.*

[6] *M. Elgendi, B. Eskofier, and D. Abbott, "Fast t wave detection calibrated by clinical knowledge with annotation of p and t waves," Sensors, vol. 15, no. 7, pp. 17 693–17 714, 2015.*

[7] *A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E.Stanley, "Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, 2000.*

[8] *E. Burns, "Ecg library," 2019. [Online]. Available: https://litfl.com/ t-wave-ecg-library/*

[9] *J. A. Vila, Y. Gang, J. M. R. Presedo, M. Fern´andez-Delgado, S. Barro, and M. Malik, "A new approach for tu complex characterization," IEEE Transactions on Biomedical Engineering, vol. 47, no. 6, pp. 764–772,2000.*