Analysis of arrhythmia features based on LightGBM and kmeans feature extraction

Zhening Liu, Zhixuan Tian, Yijun Zhou, Hao Li, Baorui Dong

Queen Mary University of London Engineering School, Northwestern Polytechnical University 710072, China

Abstract. The risk classification based on heart disease signals aims to classify patients into different risk levels, such as low risk, medium risk, and high risk. This classification can help doctors determine appropriate treatment plans, take timely measures, and improve patients' survival rate and quality of life. This article is based on the background of this problem, extracting electrocardiogram signals, and using LightGBM and Kmeans algorithms for arrhythmia classification and analysis. Extracted from the 2s features collected in this article are the data features of the electrocardiogram, including P-wave, P-R band, QRS complex, S-T band, and T-wave features. Construct a feature index pool for each electrocardiogram data, and use the feature index pool and the assigned risk level to construct a LightGBM classification regression model. The prediction accuracy is 93.5%, and the error fluctuation does not exceed ± 1. After that, use principal component analysis (PCA) to perform feature dimensionality reduction on the data. Specific classification of different categories of heart rates was carried out, and the specific performance of the algorithm and visualization of classification images were performed. Finally, this article conducted model validation on the above model, and constructed training and validation sets to validate the model. Through observation, it can be seen that the accuracy of the model is stable at over 90%. Therefore, the model constructed in this article is true, effective, reliable, and accurate.

Key words: LightGBM; Kmeans; ECG characteristics; Arrhythmias.

1. Introduction

1.1 Background Research

The widely used method for detecting arrhythmia is electrocardiography (ECG) diagnosis. Arrhythmias are an extremely common and important abnormal state of cardiac electrical activity. Different types of arrhythmia exhibit different states in the waveform and frequency of the electrocardiogram. Studying electrocardiogram signals can diagnose various symptoms of arrhythmia. Taking effective measures to detect the occurrence of arrhythmia has a great positive effect on the prevention and diagnosis of cardiovascular diseases, and has very important clinical significance. The pacemaker of the heart discharges electricity to transmit current to each myocardial fiber. Upon receiving an electrical signal, the corresponding myocardial fiber completes a contraction, and the heart beats accordingly. The electrical signals of the heart can be transmitted to the skin on the body surface, and the performance of the electrical signals detected in different body surface areas varies. In this way, electrodes are placed in specific areas of the body surface, and electrocardiogram data can be recorded through an electrocardiogram machine. For people with severe heart disease, real-time monitoring of electrocardiogram is an important means of detecting arrhythmia.

2. A Heart Classification and Recognition Model Based on LightGBM

2.1 Data preprocessing

Then, in order to ensure the accuracy of the data and the efficiency of the program, this paper conducts data processing on the power Spectral density of the ECG waveform, and carries out Fourier transform processing on the data to screen the characteristics. The specific algorithm formula is as follows :

$$f(t)=\frac{a_0}{2}+\sum_{n=1}^{+\infty}\left(a_n\cos nwt+b_n\sin nwt\right) \tag{1}$$

According to Euler's formula:

$$e^{i\theta}=\cos\theta+i\sin\theta \tag{2}$$

Fourier series can be expressed as:

$$f(t)=c_0+\sum_{n=1}^{+\infty}\left(c_n e^{inwt}+c_{-n}e^{-inwt}\right) \tag{3}$$

Ultimately, we can obtain:

$$f(t)=\frac{T}{2\pi}\sum_{n=-\infty}^{+\infty}c_n e^{in\Delta wt}\Delta w=\frac{1}{2\pi}\int_{-\infty}^{+\infty}F(w)e^{iwt}dw \tag{4}$$

By filtering, some useless noise can be filtered out and a relatively smooth waveform can be obtained for analysis. Therefore, this article uses Python software to program and filter the electrocardiogram signal, and the results are as follows:

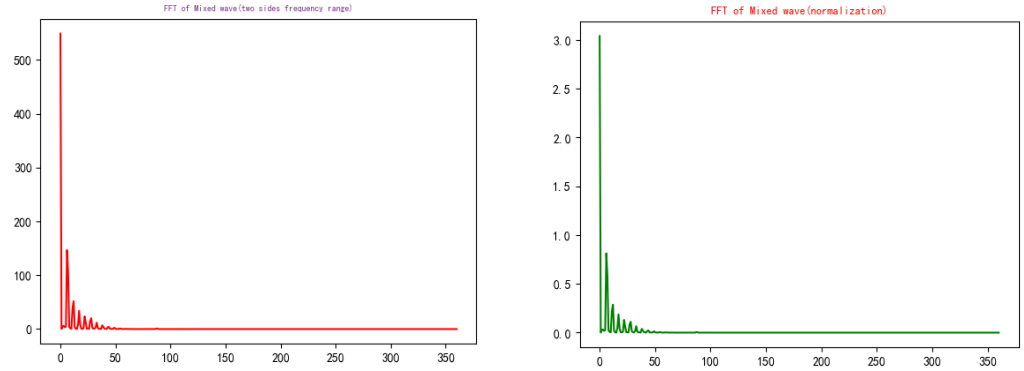

Figure 1. Filter result graph

Based on the above filtering process, we have obtained satisfactory filtering results.

2.2 A Heart Disease Classification Model Based on LightGBM

Extract the data features of the electrocardiogram from the 2s features collected in this article, and construct a feature index pool for each electrocardiogram data. Using the feature index pool and the assigned risk level, construct a LightGBM classification regression model based on decision tree improvement to solve the above problems. And by predicting the obtained label and real_ Calculate the accuracy of the model in identifying and classifying different categories based on the error size between labels.

Based on this, this article first introduces the LightGBM classification regression model under decision tree improvement, as shown below

The GBDT (Gradient Lifting Tree) model is a decision tree based model that uses a forward distribution algorithm with the idea of continuously fitting residuals to minimize them.

The negative gradient of the Loss function of the ith sample in the t round can be expressed as

$$r_{ti}=-\left[\frac{\partial L(y_i,f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{t-1(x)}} \tag{5}$$

For the samples in each leaf node, we find the output value that minimizes the Loss function:

$$C_{ti}=\arg\min_c\sum_{x_i\in R_{ti}}L(y_i,f_{t-1}(x_i)+c) \tag{6}$$

In this way, we obtain the decision tree fitting function for this round:

$$h_t(x) = \sum_{j=1}^{J} c_{tj} I(x \in Rt_j) \tag{7}$$

The final expression of the strong learner obtained in this round is as follows:

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj} I(x \in Rt_j) \tag{8}$$

Based on the analysis of the above issues, this article will use Jupyter Notebook software under Anaconda for programming and solving. The model parameters will be set as follows

Table 1. LightGBM parameter table

Parameter	Explain
squarederror	Mean square deviation as a loss function
eval_metric	MAE as an evaluation indicator
max_depth	The maximum tree depth of the decision tree is 5
learning_rate	Learning rate is 0.1
n_estimators	1000 iterations

Based on this, this article analyzes each category and calculates the difference between the predicted danger level and the actual danger level. Using MAE as the loss function, iteratively optimize the model, and the results are shown in the following figure

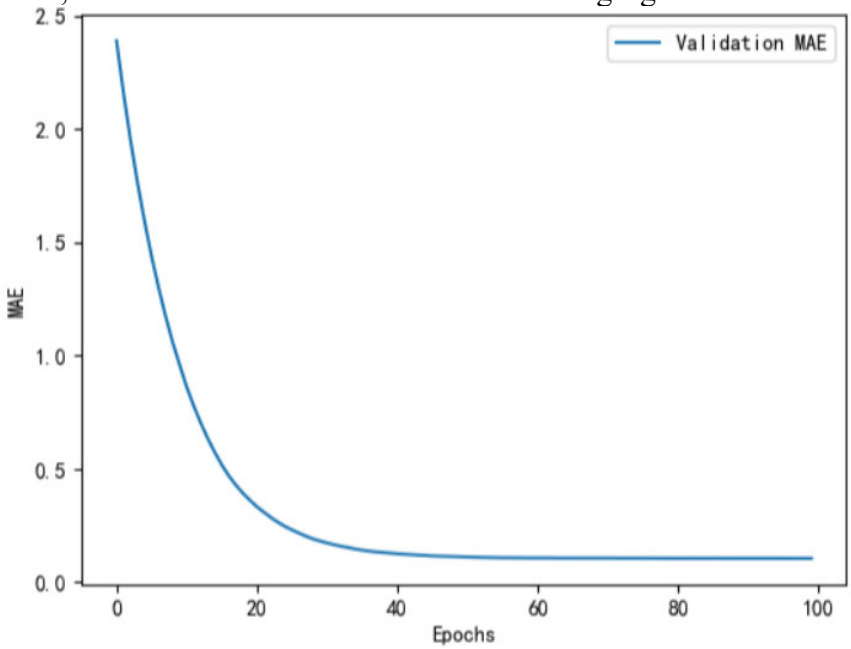

Figure 2. MAE loss graph

According to the analysis of the above figure, it can be seen that the model converges after 40 iterations and gradually stabilizes after 60 generations. Therefore, we can conclude that the above model begins to stabilize at 60 iterations. Therefore, further predictions were made for samples of all hazard levels, and the results are as follows

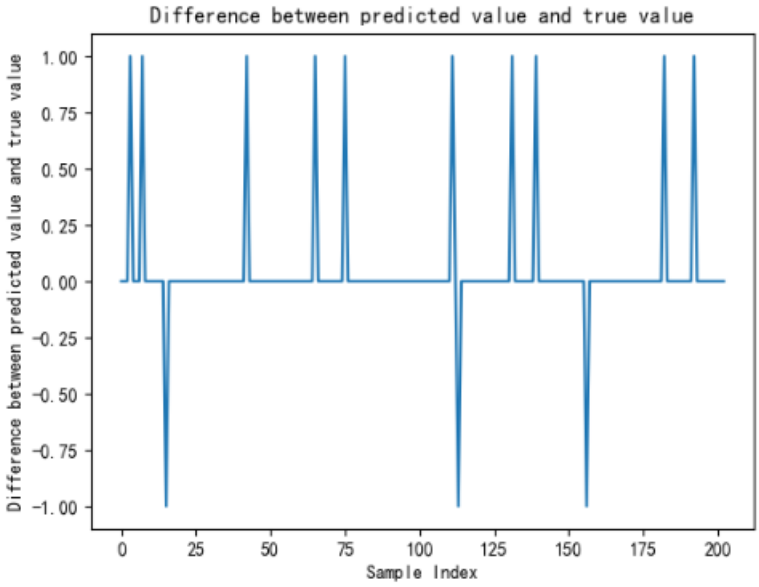

Figure 3. Predicted Difference Graph

Based on this, this article analyzes the above images and shows that a difference of 1 between the predicted value and the true value indicates a slight misjudgment, and a difference of -1 between the predicted value and the true value indicates a slight false alarm. According to the observation in the above figure, it can be seen that there have been several minor false alarms and minor misjudgments in the prediction model, but the overall prediction effect is excellent.

According to the observation in the above figure, the prediction accuracy is as high as 93.59%. This article effectively extracts features of various risk levels based on electrocardiogram morphology, and the obtained features can effectively represent the differences between patients with different risk levels. Therefore, the results obtained by regression analysis using LightGBM are true, accurate, and highly reliable.

3. Electrocardiogram signal clustering model based on km algorithm

3.1 Cluster analysis of electrocardiogram category 1 based on kmeans algorithm

We further subdivide the categories in the specific danger degree, which belongs to the problem of unknown classification and Unsupervised learning. Therefore, for the Unsupervised learning problem, this paper first uses the principal component analysis (PCA) algorithm to reduce the dimension of data features. On this basis, the elbow method is used to draw the loss value iteration line of the kmeans algorithm, and the size of the k value is determined by observing the number of iterations of the loss value, in order to classify different types of heart rates.

Clustering is an unsupervised learning method that can group similar objects into the same cluster. The more similar the objects within the cluster, the better the clustering effect. There is a significant difference between clustering and classification, and classification requires a prior understanding of the criteria for the category, allowing data to be directly classified according to certain criteria. Clustering is different. For a set of data, there is no connection between them. You can try to group objects with higher similarity into the same cluster based on their similarity.

The algorithm principle of k-means is not complex. The entire algorithm is a process of repeatedly moving the center point of a class. Firstly, k objects are randomly selected from n data objects as the initial clustering center. Through continuous iteration, the center point of the class is moved to the average position of the included members, and then its internal members are re divided until all members are closest to the center of the class they belong to:

$$\operatorname{argmin}\sum_{i=1}^{k}\sum_{x\in S_i}\left\|x-\mu_i\right\|^2=\operatorname{argmin}\sum_{i=1}^{k}\left|S_i\right|\operatorname{Var}S_i \tag{9}$$

Based on this, this article uses the Kmeans algorithm to construct kmeans clustering analysis for the data of six different categories of heart disease types in the above problems. The elbow method is used to observe the changes in the loss function, and the optimal k-value is determined for clustering analysis of different categories.

Firstly, based on the analysis of the attachment data provided by the question, it can be concluded that there are a total of 6 levels of danger. Under the premise of further subdividing them, the value of clustering number k is determined comprehensively based on the categories given in the question and the corresponding elbow graph in the KMENS algorithm.

Through understanding the problem, it can be concluded that there are no multiple categories of life-threatening arrhythmias or life-threatening ventricular arrhythmias, and this article will not further analyze them. So, first, perform cluster analysis on life-threatening arrhythmias with a risk coefficient of 1, and the results are as follows

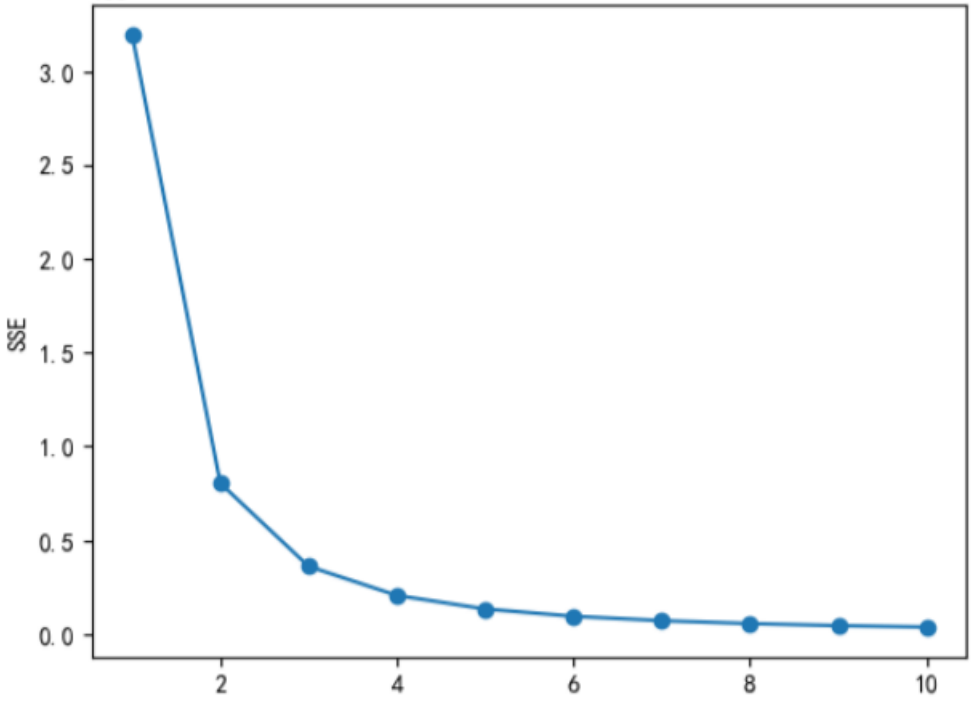

Figure 4. Further Subdivision of Hazard Category 1 by Elbow Method

Based on the above analysis, further observation of the cluster center shows that risk category I can be further divided into two categories, namely, ventricular flutter (VFL) and Ventricular fibrillation (VF).

Based on the above summary, it can be seen that the differences between these two types of anomalies are mainly reflected in the QRS and S-T, T regions. This article drew images to observe the numerical variance of the hazard level of 1 in the three regions, and the results are as follows

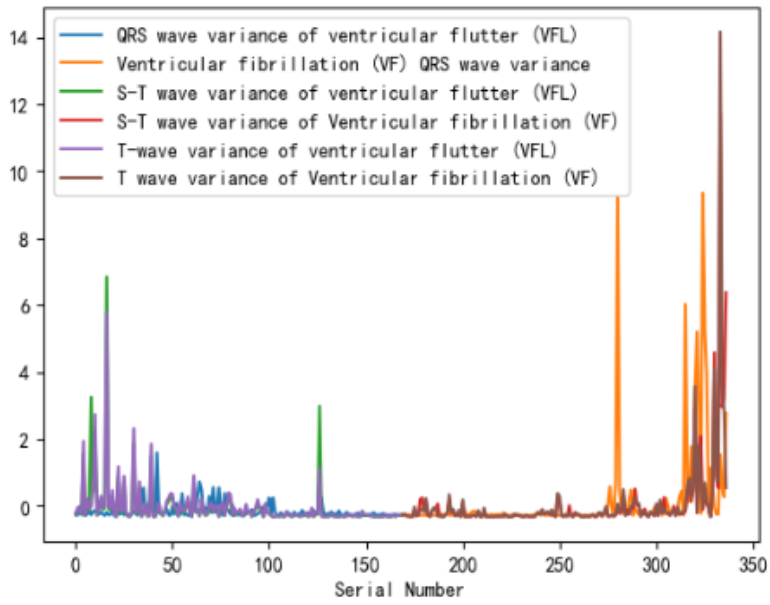

Figure 5. Run chart of two situations with hazard level of 1

By observing the differences between the two categories and analyzing them, it can be concluded that they are in line with the conclusions drawn in this article. As shown below

Ventricular Flutter (VFL) and Ventricular fibrillation (VF) are two kinds of arrhythmias. They have some differences in P wave, P-R wave band, QRS wave group, S-T wave band and T-wave images of ECG signals. The following are the main differences between them:

1. P-wave: In ventricular flutter, P-wave is usually not visible or obvious. In Ventricular fibrillation, P wave is invisible or ignored.

2. P-R band: In ventricular flutter, due to rapid ventricular activation, the P-R band may be missing or not obvious. In Ventricular fibrillation, the P-R band is usually invisible due to disordered ventricular activation.

3. QRS complex: In ventricular flutter, the QRS complex is usually a normal shape with regular waveforms. However, in Ventricular fibrillation, QRS complex showed highly irregular shape without obvious waveform regularity.

4. S-T band: In ventricular flutter, the S-T band is usually flat or slightly elevated, indicating a certain degree of ST segment elevation. In Ventricular fibrillation, the shape of S-T band is also highly irregular, which may show wide fluctuation or high variation.

5. T-wave: In ventricular flutter, the T-wave is usually normal or slightly flattened. In Ventricular fibrillation, the T wave is highly irregular, and may appear distorted, inverted, amplitude change or disappear.

3.2 Cluster analysis of electrocardiogram category 4 based on kmeans algorithm

By understanding the problem, it can be concluded that there are potential dangerous ventricular arrhythmias: low-frequency ventricular tachycardia (VTLR), ventricular premature beats (B), high ventricular ectopic beats (HGEA), and ventricular escape rhythm (VER). This article will not further analyze it. So, first, perform cluster analysis on life-threatening arrhythmias with a risk coefficient of 1, and the results are as follows

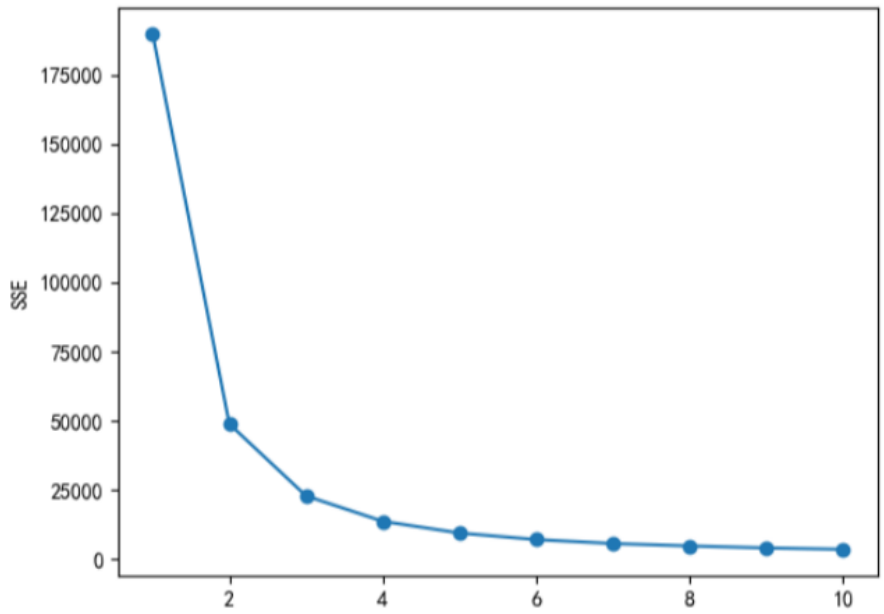

Figure 6. Further Subdivision of Hazard Category 4 by Elbow Method

Based on the above analysis and further observation of the cluster center, it can be concluded that risk category 1 can be further divided into two categories, namely low-frequency ventricular tachycardia (VTLR), ventricular premature beats (B), high ventricular ectopic beats (HGEA), and ventricular escape rhythm (VER)

Based on the above summary, this article draws images to observe the numerical variance of each area with a danger level of 4, and the results are as follows

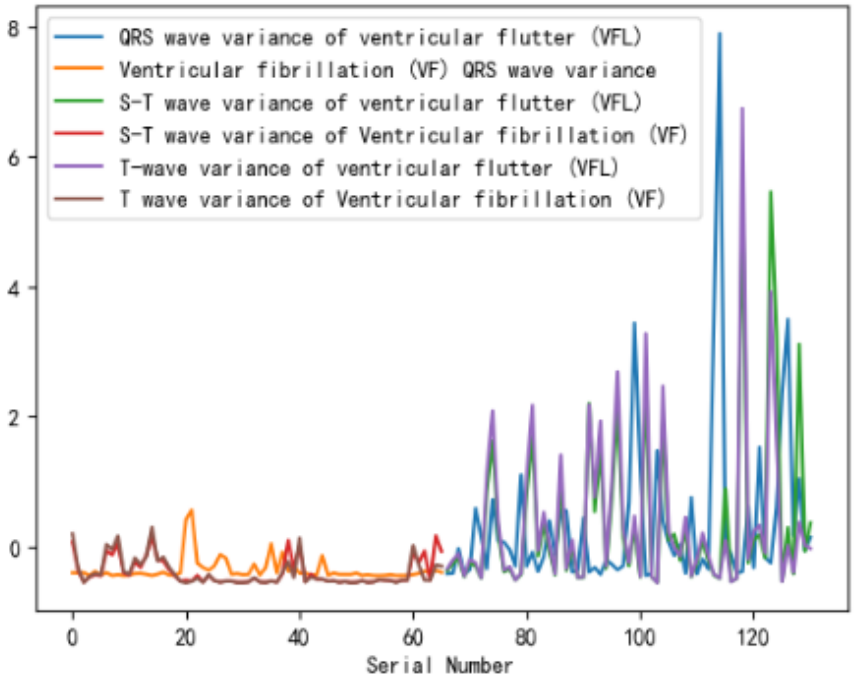

Figure 7. Run chart of two situations with hazard level of 4

Ventricular Tachycardia with Long Runs (VTLR), Couplet (B), High Grade Ectopic Beat (HGEA), and Ventricular Escape Rhythm (VER) are common types of arrhythmia. They have some differences

in the P-wave, P-R band, QRS complex, S-T band, and T-wave images on the electrocardiogram signal. The following are the main differences between them:

1. P-wave: In low-frequency ventricular tachycardia (VTLR), P-wave is usually not visible or may be masked in the QRS complex. In ventricular premature beats (B), high ventricular ectopic beats (HGEA), and ventricular escape rhythms (VER), the P-wave usually does not appear at the expected time and may exhibit abnormal morphology.

2. P-R band: The length of the P-R band may vary in low-frequency ventricular tachycardia (VTLR), ventricular premature beats (B), high ventricular ectopic beats (HGEA), and ventricular escape rhythm (VER). In low-frequency ventricular tachycardia (VTLR), the P-R band may be normal or slightly prolonged. In ventricular premature beats (B) and highly ventricular ectopic beats (HGEA), the P-R band may be shorter. In ventricular escape rhythm (VER), the length of the P-R band is usually longer.

3. QRS complex: In low-frequency ventricular tachycardia (VTLR), the QRS complex may present as a broad anomaly. In ventricular premature beats (B) and high ventricular ectopic beats (HGEA), the QRS complex may exhibit abnormal morphology, usually with a normal width. In ventricular escape rhythm (VER), the QRS complex is usually normal or slightly wider.

4. S-T band: The shape and position of the S-T band may vary in low-frequency ventricular tachycardia (VTLR), ventricular premature beats (B), high ventricular ectopic beats (HGEA), and ventricular escape rhythm (VER). In low-frequency ventricular tachycardia (VTLR), the S-T band usually exhibits a certain degree of depression or inversion. In ventricular premature beats (B) and high ventricular ectopic beats (HGEA), the S-T band may be normal or slightly elevated. In ventricular escape rhythm (VER), the S-T band is usually normal or slightly elevated.

5. T-wave: In low-frequency ventricular tachycardia (VTLR), ventricular premature beats (B), high ventricular ectopic beats (HGEA), and ventricular escape rhythm (VER), the morphology of T-wave may change. In low-frequency ventricular tachycardia (VTLR), T waves may be normal or flattened. In ventricular premature beats (B) and high ventricular ectopic beats (HGEA), T waves are usually normal. In ventricular escape rhythm (VER), T waves are usually normal or slightly flattened.

4. Conclusion

Every beat of the heart is accompanied by fluctuations in the electrical signals of the heart, reflecting whether there are problems with the cardiovascular and cerebrovascular systems. With the continuous development of technology and the continuous improvement of economic level, people often have met their material needs and started to worry about their physical health issues. Cardio cerebral Vascular disease has been the main cause of death of non communicable diseases worldwide, and has been widely concerned by the medical community. The risk classification based on heart disease signals aims to classify patients into different risk levels, such as low risk, medium risk, and high risk. This classification can help doctors determine appropriate treatment plans, take timely measures, and improve patients' survival rate and quality of life. In addition, this classification also helps to optimize the allocation of medical resources, enabling high-risk patients to receive more urgent treatment and monitoring. This article is based on the background of this issue, using the data provided by the organizer to extract electrocardiogram signals, and using LightGBM and Kmeans algorithms for arrhythmia classification and analysis.

References

[1] Fan Chengzhu Research and Implementation of ECG Automatic Classification Method Based on Deep Neural Network [D]. Shandong University, 2016.

[2] Li Kunyang, Hu Guangshu. Classification of arrhythmias based on electrocardiogram analysis [J]. Journal of Tsinghua University (Natural Science Edition), 2009,49 (03): 416-418+423. DOI: 10.16511/j.cnki. qhdxxb. 2009.03.024.

[3] Jin Linpeng, Dong Jun. Deep learning algorithm for clinical ECG analysis [J]. Science of China: Information science, 2015,45 (03): 398-416.

[4] Luo Dehan, Xu Guanggui, Zou Yuhua, et al. Research on ECG signal diagnosis model based on multi-stage artificial neural networks [J]. Journal of Instrumentation, 2008 (01): 27-32. DOI: 10.19650/j.cnki.cjsi.2008.01.006.

[5] Cao Xiwu, Deng Qinkai. Frequency analysis of ECG waves [J]. Chinese Journal of Medical Physics, 2001 (01): 46-48.

[6] Ji Hu Research on Key Technologies for Automatic Analysis of ECG Signals [D]. National University of Defense Science and Technology, 2006.

[7] Zhang Jiawei Recognition of ECG morphological features and its role in classification [D]. East China Normal University, 2011.