Research on Arrhythmia Classification Method Using Optimized Probabilistic Neural Network

Liu Junkong\textsuperscript{1,a}, Li Lili\textsuperscript{1,b*}, Fang Yipin\textsuperscript{1}, Su Jinyu\textsuperscript{1}

\textsuperscript{1}State Key Laboratory of Mechanical Manufacturing Systems Engineering, Xi’an Jiaotong University, Xi’an, Shanxi, China
\textsuperscript{a}email: i1196689756@stu.xjtu.edu.cn, \textsuperscript{b}email: lilili199013@stu.xjtu.edu.cn

Abstract: Aiming at the classification difficulty of complex and diverse ECG signals, this paper proposes a feature extraction method based on standard deviation. This method solves the problems of multi-dimensional, diverse, high similarity, and the difficulty in extracting main features effectively of ECG(electrocardiogram, ECG) signal features. In addition, this method overcomes the difficulty of low classification accuracy because of large differences in ECG signals of the same type among different patients. This paper adopts optimized probabilistic neural network methods to achieve automatic classification of arrhythmia. First, the standard deviation of the dimensions of each sampling point of the ECG signal would be calculated and sorted by size. Second, the first m dimensions as the feature dimensions of arrhythmia would be extracted. After that, the probabilistic neural network would be used to train and classify feature dimension data. Finally, Bayesian Optimization (BO) method would be used to optimize the parameters globally. In the experiment of the MIT-BIH arrhythmia database, the arrhythmia data was divided into 5 categories and verified, experimental results show that the correct rate of classification of the arrhythmia data of patients reached 99.67%, which proved the effectiveness of the method in this paper.

1. Introduction
An unhealthy diet, insufficient physical exercise, excessive drinking, smoking, and other living habits can lead to cardiovascular disease\cite{1}. Arrhythmia is an important manifestation of cardiovascular disease. The diversity of arrhythmias can be shown by ECG. ECG contains abundant arrhythmia information. Clinically, ECG is very common and effective as an auxiliary diagnostic method. ECG analysis by experts not only has the defects of misjudgment and subjectivity, but also consumes a lot of energy, and even causes serious consequences such as delay of illness. In daily life, intelligent wear is becoming more and more popular. Smart wear is becoming more and more popular. It is very common to detect the wearer's ECG through smart wear. It will become a reality to analyze the ECG and then provide health advice to the wearer. Therefore, the automatic classification of ECG is of great significance.

In recent years, the research of arrhythmia mainly includes signal denoising, feature extraction and dimensionality reduction, automatic classification, and so on. The purpose of signal processing is to remove interferences. Such as, power frequency noise and EMG signals accompanied the ECG signal, the signal was segmented at the same time to facilitate feature extraction and automatic classification. For example, Yin Li et al. proposed an ECG adaptive denoising method based on ensemble empirical mode decomposition (EEMD) improved threshold function for ECG signals with baseline drift, power frequency, and EMG interference, which is complete on the basis of significant denoising. It retains
the waveform characteristics[2], although the effect is significant, the process is cumbersome. Although the effect is significant, the process is cumbersome. Zheng Minmin et al. used sym5 wavelet function to decompose ECG signal by 8-layer wavelet and obtained a new threshold function, and then used the proposed new threshold function for wavelet threshold de-noising[3], which is difficult to select the wavelet basis. ECG feature extraction is mainly based on principal component analysis and other feature dimensionality reduction methods to obtain key features. For example, Ye et al. combined wavelet transform and independent component analysis with ECG morphological features to obtain an effective ECG feature set[4]. Although effective, it has certain blindness. Machine learning is widely used in ECG automatic classification. YIN et al. used a multi-domain ECG feature extraction method for the accurate classification of 8 types of arrhythmia, with a classification accuracy of 99.70%[5], but there is a lack of research on the predictive classification of ECG. MENG et al. used the original ECG signal and the RR interval as input, and achieved 6 types of arrhythmia classification through 4-layer deep belief networks (Deep Belief Networks, DBNs), with an overall accuracy of 98.49%[6], and there are fewer types of arrhythmias. Li et al. used the least squares support vector machine pair based on a genetic algorithm to realize the classification of five arrhythmias, and achieved 99.14% classification accuracy[7]. Song Lixin et al. used discriminant-based ddbns, with an accuracy of 99.31%[8]. Keli et al. used the improved long-short memory network (LSTM) method to achieve 99.2% classification accuracy[9]. Xiong Hui et al. used a 7-layer hybrid model structure with a convolutional neural network (CNN) as the core to achieve four types of arrhythmias, with a classification accuracy of 99.16%[10]. The above ECG classification has achieved remarkable results, but there are still some limitations: ① The data preprocessing is cumbersome; ② The ECG waveforms between patients have great differences, and the feature extraction is difficult; ③ There are few studies on arrhythmia types with fewer data. In view of the above shortcomings, this paper proposes an optimized probabilistic neural network ECG classification method: firstly, the sampling points are sorted according to the standard deviation of the same kind of heartbeat sampling points, and the dimension with a smaller standard deviation is extracted as the feature of this kind of heartbeat, then the PNN model is established, and finally, the Bayesian Optimization (BO) is used to optimize the extraction feature proportion C and the smoothing factor δ of PNN, so as to achieve the goal of ECG classification purpose.

2.Data and methods
The flow of the automatic classification algorithm for arrhythmia is shown in Figure 1. The data is preprocessed first. The main work of preprocessing is: denoising the ECG signal, then locate the peak, and intercept the heartbeat based on the peak.

Then comes the feature extraction part. A single heartbeat contains 260 sampling points, which has more dimensions and contains a large amount of information. Feature extraction has a direct impact on the following classification and prediction effects. Feature extraction can be divided into two categories: ① decompose the wave and extract features[11]; ② principal component analysis. These two methods are to preserve all the features of the signal as much as possible, and principal component analysis can reduce the dimension of the data. The disadvantage of these two feature extraction methods is that they contain almost all the features of the data so that the main features of the data are not obvious, and the secondary features become interference features. This paper proposes a new feature extraction method, the feature extraction method based on standard deviation. This paper proposes a new feature extraction method, a feature extraction method based on standard deviation. This method can not only extract the main features of the data, but also discard the secondary features, and reduce the data dimension at the same time. The method first calculates the standard deviation of each dimension of the data, and the calculation formula of standard deviation s is as follows. The small standard deviation indicates that the data fluctuation of this dimension is small, and the characteristics of each data are the main characteristics or fundamental characteristics. Therefore, the standard deviation of each dimension is sorted according to the size, and the top m dimension is taken as the characteristic data of the data.
\[ x = \frac{x_1 + x_2 + \cdots + x_n}{n} \]  
\[ s = \sqrt{\frac{(x_1 - x)^2 + (x_2 - x)^2 + \cdots + (x_n - x)^2}{n}} \]

(1)

(2)

Because PNN has a super parameter smoothing factor \( \delta \), the above feature extraction dimension is also a super parameter. In order to achieve the highest classification accuracy, this paper uses BO to globally optimize the dimension of feature data and the super parameter smoothing factor \( \delta \) of PNN, and finally completes the training of PNN and makes classification prediction.

2.1. MIT-BIH arrhythmia database

In this paper, the experimental data of arrhythmia classification is divided into "between patients". The training set and test set are divided according to the scheme proposed by Chazal[12]. The ECG records in the MIT-BIH database are divided into two groups: DS1 (training set) and DS2 (test set). The training set DS1 includes the following 22 ECG records of MIT-BIH: 101, 102, 106, 107, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223 and 230; the test set DS2 includes The 24 ECG records were 100, 103, 104, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 217, 219, 221, 222, 228, 231, 232, 233 and 234. Table 1 shows the number of beats for five arrhythmia types.

| Types of arrhythmia | N   | V   | S   | F   | U   | Total |
|---------------------|-----|-----|-----|-----|-----|-------|
| Types of heart beat | N,L,R,e,j | V,E | A,a,J,S | F | /,f,Q |       |
| Training set DS1    | 45401 | 3081 | 935 | 409 | 2078 | 51904 |
| Test set DS2        | 44482 | 2495 | 1803 | 379 | 1809 | 50968 |
| Total (DS1+DS2)     | 89883 | 5576 | 2738 | 788 | 3887 | 102872 |

2.2. Data preprocessing

In this paper, "wfdb2mat" in the open-source software package WFDB is used to correct and process the baseline data, and the gqrs algorithm is used to relocate the R wave peak before the heartbeat is intercepted. 95 sampling points on the left side of the R wave peak and 165 sampling points on the right side (including the R wave peak) are taken as a heartbeat. Figure 2 shows the single beat waveform randomly selected from each type of arrhythmia after pretreatment.
2.3. Feature extraction

The advantage of the standard deviation feature extraction method proposed in this paper is that it reduces the dimensions of the data, facilitates data processing, and at the same time has a screening effect on features, and only retains the main features of the data. After preprocessing the data, we get five types of beat data. The data of each beat type contains 260 sampling points, which are recorded as 260 dimensions. It can be seen from Table 1 that class N cardiac beat contains five types of arrhythmia (n, l, R, e, J), with a total of 45401 cardiac beat training sets. Class N cardiac beat has many types of arrhythmia and contains arrhythmia cardiac beat of different patients. There are great differences in cardiac beat, and the common characteristics are not obvious. The steps of the standard deviation feature extraction method are as follows: first, calculate the standard deviation of the first dimension (sampling point) of 45401 heartbeats, similarly, count the nth (n = 2 The standard deviation of dimension data is $S_n[13]$, and then $S_n$ is sorted from small to large to get the standard deviation sequence L. Then the dimensions of the first m data in the L sequence are called effective dimensions, and the rest are invalid dimensions. The data of invalid dimensions are deleted. From the previous analysis, it can be concluded that the higher the dimension of the data in the standard deviation series L, the smaller the data fluctuation, the more important the data of this dimension, and the more representative of this type of characteristics. All types of data of training set and test are extracted according to the above method to prepare for classification.

Extract 100 valid dimension data from five types of data in the training set to get a single beat waveform, as shown in Figure 3.
2.4. Classification and evaluation

The network structure of Probabilistic Neural Network (PNN) is similar to that of RBF neural network, but the difference is that PNN is a forward propagation network and does not need backpropagation to optimize parameters. This is because PNN combines Bayesian decision-making to determine the category of test samples. Suppose that there are t possible categories for test sample T: \{w_1, w_2, \ldots, w_t\}, then the Bayesian decision to judge the sample category is \( \max \{ p(w_1 \mid T), p(w_2 \mid T), \ldots, p(w_t \mid T) \} \).

PNN network structure is divided into four layers: input layer, mode layer, summation layer, and output layer. Suppose the training sample is \{t_1, t_2, \ldots, t_L\}, and the number of samples is L. The interaction of PNN layers and the relationship between them are described as follows.

**Input layer:** input test samples, the number of nodes is equal to the sample feature dimension.

**Pattern layer:** calculate the Gauss function value of each sample in the test sample and training sample, and the number of nodes is equal to the number of training samples. The value of Gauss function between the test sample \( t \) and the jth training sample \( T_j \) (for the test sample \( x \), the value output from the jth mode layer node) is:

\[
Gauss(T - t_j) = e^{-\frac{||T_j - x||^2}{2\sigma^2}}
\]

Where \( \delta \) is the super parameter of the model.

**Summation layer:** the output sum of the pattern layer nodes corresponding to the same class of test samples is obtained, and the number of nodes is equal to the number of training samples. Output layer: normalize the output of the above summation layer, calculate the probabilities of different categories of test samples, and judge the category of test samples according to the probability, with the number of nodes being 1.

The m-dimension training data obtained above is used for the PNN, and the radial basis function (Gauss function) is used to classify the test data in PNN model layer. All the classification results were divided into four types: TN (true negative), FN (false negative), TP (true positive), FP (false positive). The following three evaluation indexes were used to evaluate the classification effect, namely classification accuracy (ACC), specificity (SPE), sensitivity (SEN)[5].

\[
Acc = \frac{TP}{(TP + FP)} \times 100\% \tag{4}
\]

\[
Spe = \frac{TN}{(TN + FP)} \times 100\% \tag{5}
\]
3. Results & Discussion

According to the method of feature dimension extraction based on the standard deviation mentioned above, feature extraction is carried out for training data and test data, and the two parts of data keep the same dimension. Then PNN is trained and classified. The selection rules of arrhythmia training data are: N, V, U randomly selected 1000 heartbeat data. S, F all selected as training data, a total of 4344 training data, test data is 50968. This method has two super parameters: the dimension proportion of data C (the percentage of extracted feature dimension in the total dimension) and the smoothing factor $\delta$ of PNN. The value range of $C$ is set to (0.2,0.4), and the value range of $\delta$ is set to (0.01,0.1). The global optimization algorithm Bayesian Optimization is used for optimization[14]. Figure 4 shows the result of 20 iterations, and the accuracy is 99.67% ($C = 0.4, \delta = 0.02$).

In addition to the overall classification accuracy, the classification results of each type of heartbeat are also very important. Table 2 shows the values of the above three indicators of each beat when the classification result is 99.67%. As can be seen from the histogram of classification results in Fig. 6, the classification indexes of five types of cardiac beats are more than 95%, especially the N, F, and U classification indexes are more than 99%.

Moreover, the number of samples of class F is 409, which is much smaller than the number of samples of other types (the number of samples of class N is 45401, the number of samples of class V is 3081, the number of samples of class S is 935, and the number of samples of U is 2078), however, the recognition rate of class F is 99.5%, and the recognition rate is very high, indicating that the method proposed in this paper also has a high recognition rate for small sample types of ECG signals.

Consequently, the classification results of five kinds of heart beat in Table 2 show that the effect of automatic classification of arrhythmia by the optimized probabilistic neural network is remarkable.

Table 2 classification results of five kinds of heartbeat

| Index | N     | V     | S     | F     | U     |
|-------|-------|-------|-------|-------|-------|
| Acc(%)| 99.9662 | 96.5638 | 96.6186 | 99.4737 | 99.8885 |
| Sen(%)| 99.8494 | 99.1182 | 96.6722 | 99.7361 | 99.0050 |
| Spe(%)| 99.7656 | 99.8182 | 99.8758 | 99.9960 | 99.9959 |

Table 3 summarizes other methods of data classification, which are based on the MIT-BIH database. It can be seen from the table that the classification method of arrhythmia proposed in this paper has high accuracy.

Table 3 performance comparison of arrhythmia classification methods

| Research | Type | Classifier | Performance (%) |
|----------|------|------------|-----------------|
|          |      |            | Acc  | Sen  | Spe  |
|          |      |            |      |      |      |
4. Conclusions
This paper proposes a feature extraction method based on standard deviation: the feature dimensions were sorted according to the size of the standard deviation of the feature dimensions, and then the first \( m \) dimensions with the smaller standard deviation were exacted as the effective dimensions. This method can not only extract the main features of the data from the multi-dimensional data, discard the secondary features, the secondary features from interfering with the classification results of ECG signals would be prevented, but also the dimensionality of the data would be reduced. At the same time, the preprocessing requirements of the data were not high. The method proposed in this paper is verified with the data of the MIT-BIH database, and the extracted features are trained and recognized by Bayesian Optimization-Probabilistic Neural Network, the classification correct rate reached 99.67\%, which proves the effectiveness of the method of this paper.

In addition, the recognition and classification results of class F ECG signals show that the method proposed in this article is also suitable for accurate recognition of small sample types of ECG signals.

Acknowledgements
The authors gratefully acknowledge the financial supports by the National Key Research and Development Program of China under Grant No. 2019YFB1703800. The authors also sincerely thank the reviewers for their detailed recommendations and comments.

Reference
[1] Anand, s., Salim, Y. (2011) Stemming the global tsunami of cardiovascular disease[J]. The Lancet,377(9765): 529-532.
[2] Yin, L, Chen, F, Zhang, Q, Chen, X. (2020) ECG adaptive denoising method based on ensemble empirical mode decomposition and improved threshold function. Journal of Xi'an Jiaotong University. 54(01):101-107.
[3] Zheng, M, Gao, x, Xie, H. (2017) Research on an Improved Algorithm for Wavelet Denoising of ECG. Chinese Journal of Biomedical Engineering. 36(01):114-118.
[4] Ye, C, Vijaya Kumar, B V K, Coimbra, M T. (2012) Heartbeat Classification Using Morphological and Dynamic Features of ECG Signals. IEEE T Bio-Med Eng. 59(10):2930-2941.
[5] Yin, L, Chen, F, Zhang, Q, Ma, X. (2019) Arrhythmia classification based on multi-domain feature extraction. J. Phys.: Conf. Ser. 1237:22062.
[6] Meng, H, Zhang, Y, (2014) Classification of Electrocardiogram Signals with Deep Belief Networks, IEEE International Conference on Computational Science and Engineering. Chengdu, pp. 7-12.
[7] Duan, L, Zhang, H, Zhang, M. (2016) Wavelet De-Noiseing and Genetic Algorithm-Based Least Squares Twin SVM for Classification of Arrhythmias. Circ Syst Signal Pr. 36(7):1-19.
[8] Song, L, Sun, D, Wang, Q, Wang, Y. (2019) Automatic classification of arrhythmias based on Discriminant deep confidence network. Journal of biomedical engineering. 36(03):444-452.
[9] Ke, L, Wang, D, Du, Q, Jiang, C. (2020) Classification of arrhythmias based on convolutional long-term and short-term memory network. Journal of electronics and information. 42(08):1990-1998.
[10] Xiong, H, Liang, M, Liu, J. (2021) Arrhythmia classification algorithm based on convolution neural network hybrid model. Journal of Harbin Institute of Technology. 53(02):33-39.
[11] Sun, H, Wang, X, Liu, X, Zhang, L, Zhao, X. (2021) Classification of EEG signals in frequency domain attention mechanism. Journal of Xi'an Jiaotong University. 55(02):129-135.
[12] Chazal, P D, O'Dwyer, M, Reilly, R B. (2004) Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE transactions on bio-medical engineering. 51(7):1196-1206.
[13] Zhang, J. (2003) Determination method of weight coefficient based on standard deviation and mean deviation and its application. Shu li tong ji yu guan li. 22(3):22-26.
[14] Gao, Z, Bao, T, Li, Y, Wang, Y. (2021) Dam deformation prediction model based on Bayesian optimized lightgbm. Journal of Changjiang Academy of Sciences:1-6.
[15] Sabut, S, Sahoo, S, Kanungo, B, Behera, S. (2017) Multiresolution wavelet transform based feature extraction and ECG classification to detect cardiac abnormalities. Measurement. 108(108):55-66.
[16] Tan, J H, Hagiwara, Y, Pang, W, Lim, I, Oh, S L, Adam, M, San Tan, R, Chen, M, Acharya, U R. (2018) Application of stacked convolutional and long short-term memory network for accurate identification of CAD ECG signals. Computers in Biology & Medicine:19-26.
[17] Martis, R J, Acharya, U R, Min, L C. (2013) ECG beat classification using PCA, LDA, ICA and Discrete Wavelet Transform. Biomed Signal Process. 8(5):437-448.
[18] Acharya, U R, Oh, S L, Hagiwara, Y, Tan, J H, Adam, M, Gertych, A, Tan, R S. (2017) A deep convolutional neural network model to classify heartbeats. Comput Biol Med. 89:389-396.
[19] Zhai, X, Chung, T. (2018) Automated ECG Classification using Dual Heartbeat Coupling based on Convolutional Neural Network. IEEE Access PP:1.
[20] Fatin, A., Elhaj, Naomie, Salim, Arief, R., Harris, Tan, Tian. (2016) Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals. Computer Methods & Programs in Biomedicine.