Journal of Ambient Intelligence and Humanized Computing (2022) 13:3945–3962
https://doi.org/10.1007/s12652-021-03492-3

ORIGINAL RESEARCH

RecogHypertension: early recognition of hypertension based on heart rate variability

Hongbo Ni[1] · Zhuang Li[1] · Ziqiang Shao[1] · Mingzhe Guo[1] · Junqi Liu[1]

Received: 6 May 2020 / Accepted: 6 September 2021 / Published online: 30 September 2021
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2021

Abstract

In recent years, the number of patients with hypertension is increasing, but the early symptoms of hypertension are not obvious, the incubation period is long, and the awareness rate and control rate are very low. Therefore, it is necessary to study the early recognition of hypertension in a non-clinical environment. The blood pressure of human being is controlled by autonomic nervous system, and heart rate variability (HRV) is an impact of autonomic nervous system and an indicator of the balance of cardiac sympathetic nerve and vagus nerve. So HRV is good method to recognize the hypertensive patients from healthy person. In this paper, we proposed a fined-grained HRV analysis method to recognize hypertensive patients from healthy person. Specifically, we cut the 8 h of ECG data into 5 min segments at first, and then we propose an improved heartbeat interval extraction algorithm to extract the heartbeat interval from Electrocardiogram (ECG) data and we extract 22 HRV features in linear, nonlinear domain and histogram, Specially, we model the distribution of the heartbeat interval of each time window using a Gaussian mixture model. Next we analyzed the correlation between linear domain and nonlinear domain features of heart rate variability. Finally, we use common machine learning algorithms to train a recognition model for hypertension. In this paper, we use 138 hypertension patients' and 138 healthy person real-world clinical Electrocardiogram data as our data set. The recognition precision rate for patients with hypertension is 97.1%, and the recall rate is 97.1%. The experimental results validate the effectiveness and reliability of the proposed recognition method in this research.

Keywords Heart rate variability · Heartbeat interval extraction · Gaussian mixture model · Hypertension recognition model

1 Introduction

Hypertension is a common and chronic disease. According to the survey, the number of patients with hypertension worldwide continues to rise. According to the online research published in the Lancet magazine on November 15, 2016, the number of patients with hypertension in the world in 1975 was 594 million. In 2015, this number has exceeded 1.1 billion, which means that since 1975, the number of adults with high blood pressure in the world has nearly doubled in 30 years (NCD-RisC 2016). In China, by the end of 2015, the prevalence of adults aged 18 and over reached 23%, and the number of patients was about 243.5 million. However, the early symptoms of hypertension are not obvious, and the incubation period is long. In the early years of the onset, there are no obvious symptoms and it is not easy to be detected. So it is called "silent killer" (Li et al. 2016). In 2012, The annual awareness rate is only 46.5%, and the control rate is only 13.8% in china, which means that more than half of the hypertensive patients still do not know their condition and miss the best treatment opportunity (Wang et al. 2016). Therefore, it is extremely important to study the early identification of hypertension in a non-clinical environment.

Heart rate variability (HRV) reflects the autonomic nervous system activity and quantitative evaluation of cardiac sympathetic nerve and vagus nerve tension and balance, to

✉ Zhuang Li
3198054800@qq.com

Hongbo Ni
nihb@nwpu.edu.cn

Ziqiang Shao
599761637@qq.com

Mingzhe Guo
guomingzhe96@qq.com

Junqi Liu
28354629@qq.com

[1] Northwestern Polytechnical University, Xi'an, China

determine the washing is an important indicator of cardiovascular disease. ECG is an effective measure and record the details of the electrical activity of diagnostic equipment. ECG signals come from the Electrocardiogram (ECG) is one of the most obvious features of QRS complex, contains from the ventricular electrical activation of P, Q, R, S,T and U wave(Van Oosterom 2009). The heartbeat interval refers to the distance between two adjacent R waves, which is also called the RR intervals, and Heart rate variability is a general term for all features extracted from the heartbeat interval.

Based on the above statements, it is a good method to recognize hypertensive patients from healthy people by HRV extracted from the ECG in sleep stage. Instead of extracting features with 8 h ECG data directly, we cut entire nights' ECG with 5 min segments and extract features from 5 min segments, which is a fine-grained analysis method and can enhance the precision and recall of the recognition of hypertension.

To the best of our knowledge, the problem of recognizing hypertensive patients from healthy person leveraging ECG data has not been well investigated in the literature. There are several challenging questions to be answered. How to extract the heartbeat interval from the ECG data, while avoiding the false and missed detection of R waves as much as possible? How to extract effective features to recognize the hypertensive patients from heartbeat interval sequence? The linear domain and nonlinear domain features come from the same data source. How relevant are they?

To answer these questions, we propose RecogHypertension, a system that predicts the healthy status (Hypertension or health) of the unknown person leveraging ECG data during nighttime sleep.

We first extract the ECG data for each person from 10:00 to 6:00 in the evening. Then we extract heartbeat interval from ECG data, and propose an improved heartbeat interval extraction algorithm to correct the heartbeat interval between the false detection and the missed detection. Using these data, at first, we extract linear domain and nonlinear domain features, then we model the distribution of heartbeat intervals for each segment using a Gaussian mixture model (GMM) and use Expectation Maximum (EM) Algorithm to solve unknown parameters. Finally, we train a classification model to predict the healthy status of unknown person. We make the following contributions.

1. This work proposes an improved heartbeat interval extraction algorithm to correct the heartbeat interval between the false detection and the missed detection. By observing the features of ECG signals, we propose a targeted R-wave error detection and miss detection algorithm.
2. We not only extract the linear domain and nonlinear domain features from the heartbeat interval sequence, but also model the distribution of the heartbeat interval of each segment with a Gaussian mixture model. This modeling method is applicable to the modeling of other physiological parameter data distribution.
3. To quantitatively analyze the correlation strength between linear domain and nonlinear domain features, we use Pearson correlation analysis method to calculate the correlation coefficient between each two features of each people, and combine the physiological reasons of the features to analyze the reasons for strong correlation between some features. What is more, we propose a feature selection method based on correlation strength and information gain.

2 Related work

2.1 Traditional hypertension recognition method

The most common and traditional method of diagnosis of hypertension is to determine whether you have high blood pressure by measuring blood pressure (Coccagna and Lugaresi 1978; Fletcher and Levin 1984).

One of the methods is to measure blood pressure directly, but this method requires the tester to wear an inflatable cuff, fingertip cuff, etc. Davies et al. (1994) dedicated measurement equipment, and cannot continue to measure, and the measured blood pressure value will be affected by the measurement person's personal status and the surrounding environment, resulting in inaccurate results. In view of the shortcomings of direct measurement of blood pressure in a non-clinical environment, many studies have proposed a method of indirectly measuring blood pressure based on the pulse wave conduction time, which refers to the pressure wave after the heartbeat spreads between two arterial sites. The time delay, the speed of this forward wave depends on the support of the artery. When the artery is enhanced, the speed of pulse wave propagation is accelerated. Since the arterial blood pressure is proportional to the blood pressure, the pulse conduction time is inversely proportional to the blood pressure of the person, and the pulse wave conduction is obtained. After the time, you can get blood pressure according to the calculation formula. Zheng et al. (2013) monitored the ECG and PPG signals simultaneously on the arm, and then used these signals to estimate the blood pressure during nighttime sleep based on the pulse wave conduction time; Wiens et al. (2017) used smart watches to monitor the micromotion caused by the heartbeat. Then, the pulse transit time is calculated, but this method is more accurate only when the tester stands. The above two measurement methods require wearing wearable devices, and these devices may cause measurement results may be inaccurate due to the tightness of the wearing. Carek et al.

(2017) and Carek and Holz (2018) used both acceleration and optical sensors to obtain pulse Conduction time, specifically the integration of the acceleration sensor in the shorts and contact with the wearer's torso, while the optical sensor observes the pulse wave reflection of the femoral artery on the wearer's thigh. This method requires the tester to wear a special wear during nighttime sleep. Shorts, and will be disturbed by body movements during sleep, and the limitations are relatively large.

2.2 Hypertension recognition method based on heart rate variability

In recent years, heart rate variability has been used in many studies to study the identification of hypertension.

Ni et al. (2017) used the belt to continuously collect electrocardiogram (ECG) data during the night's sleep, first dividing the data of the whole night into data segments of different time scales such as 8 h, 4 h, 2 h, 1 h, and 30 min. Constructing a time series pyramid, the time scale of the next layer of data is half of the upper layer, and then extracts the time domain, frequency domain and nonlinear domain features from the heartbeat interval of each time segment, so that each layer of The data segments correspond to a feature vector, and then, in order to reduce the dimension, a plurality of feature vectors are aggregated into one feature vector by using a pooling method in each layer, and an average value of all the layers of each feature is obtained between the layers. As the final value of the feature, and then using the feature training model, the classification accuracy rate of the two types of people can reach 93.33%. This method solves the problem that only considers a single time scale in most studies. The focus of the research is on how to The selection of features between different time scales, but this paper divides the data of the whole night into a 5-min time window, extracts features from multiple angles, and studies the heart rate variability features in fine-grained manner. A correlation and volatility model, a new feature selection method was proposed; MG Poddar et al. (2014) and others used a 5-min electrocardiogram (ECG) record collected in a clinical environment of 57 healthy people and 56 hypertensive patients. The classification accuracy of the support vector machine model using all time domain, frequency domain and nonlinear domain feature training can reach 100%. However, in this article, the author simply uses all features for classification, using only 5 min of research objects. The records were analyzed, and this paper analyzed the ECG signals for 96 consecutive periods of time, avoiding the possible impact of accidental factors on the classification results in the process of collecting ECG data, and the results were more convincing and credible. Song et al. (2015) used BCG signals collected from fretting sensitive mattresses to classify healthy people, hypertensive patients, and patients with coronary heart disease. This study proposes a new method for extracting heartbeat intervals using collective empirical mode decomposition, and then the time domain, frequency domain and nonlinear domain features were extracted from the heartbeat interval, and the differences between these three features were compared. The t-test results showed that these features had significant differences among the three types of people, and then respectively. Using the time domain, frequency domain, nonlinear domain and the combination of these features, the model is trained by Naive Bayesian algorithm. The results show that the accuracy of classification using the combination of three types of features is higher than that of using single class features. The main contribution of this paper is to propose a new method to extract the heartbeat interval. The focus of this paper is on feature extraction, feature analysis and modeling of hypertension recognition models.

3 Data acquisition and analysis

ECG is an important bioelectricity that embodies the physiological state of various parts of the human heart. In this study, we use ECG data for two types of people, patients with hypertension used the SHAREE (Smart Health for Assessing the Risk of Events via ECG) data set downloaded from the PhysioNet website (2015). The dataset includes 139 patients' 24-h Holt recording recruited from the Naples Federico II University Hospital Hypertension Center in Italy (data from 138(including 90 males), in this article, one of which was discarded due to less than 24 h) after one month of antihypertensive treatment. The data set for healthy people comes from the Telemetric and Holter ECG Warehouse (THEW) database (Couderc et al. 2005), which is run by the University of Rochester Medical Center. The dataset contains data for 202 people. For healthy people, this article uses the data of the previous 138 people (including 71 males). This paper studies the classification of hypertensive patients and healthy people by using ECG signals continuously collected by two types of people during nighttime sleep. Ni et al. (2019) studies have shown that people's physical parameters in the sleep state are more stable, less affected by the external environment, can more accurately reflect the various functions of the body, and the duration of sleep is relatively long. There are fewer activities when people in the sleep process, and the collected data is less noisy. The data can fully reflect the dynamic changes and intrinsic subtle changes of the heart rate variability of the subjects, so we use the data of hypertensive patients and healthy people from 10:00 to 6:00 the next morning. The hourly data was studied. During the study, all subjects were in the sleep state by default. After extracting the data for the specified time period, the data of 8 h in the whole night is divided into 5 min, that is to say, 96 segments, in the following analysis of data.

4 Problem statement and system framework

4.1 Problem statement

The problem can be stated as follows: given the ECG signal of a person sleeping at night, to determine whether it is a hypertensive patient, this problem is essentially a two-category problem.

The problem can be formalized as that the heartbeat interval sequence is obtained from the electrocardiogram signal of the study subject, and the feature set F is extracted from the heartbeat interval sequence, and we want to predict the category C (C = 0, 1). Let F = {F_1, F_2,…, F_n}, C = {0, 1}, given $F^{(1:t+1)}\left(=\left\{F^{(1)},\dots,F^{(t+1)}\right\}\right)$ and $C^{(1:t)}\left(=\left\{C^{(1)},\dots,C^{(t)}\right\}\right)$, our objective is to predict C(t + 1).

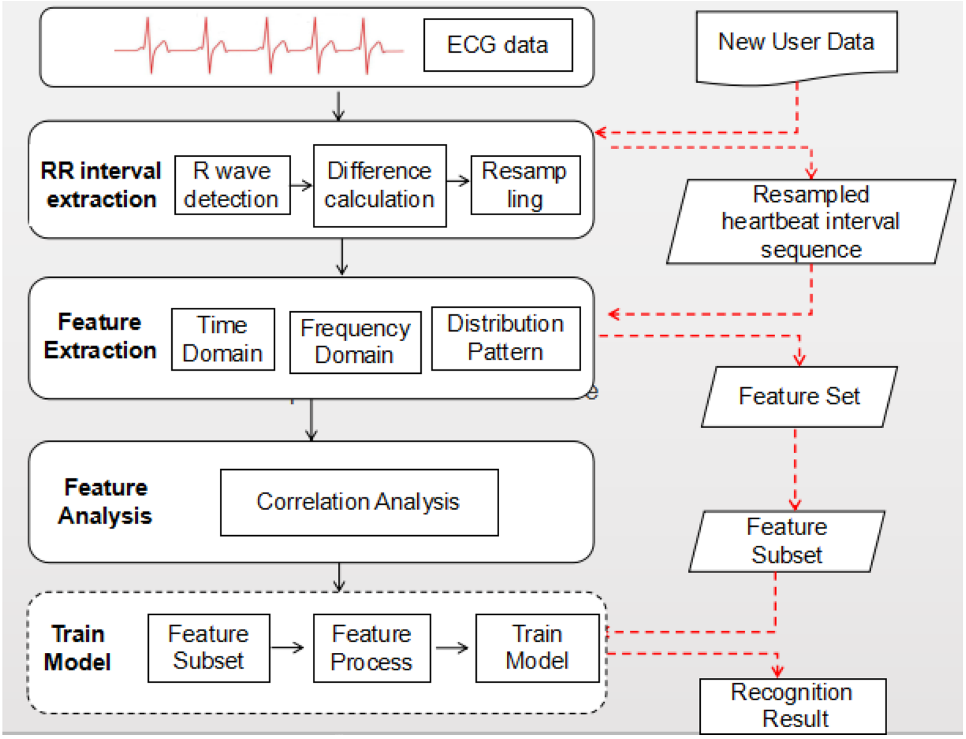
4.2 System framework

The overview of the framework is illustrated in Fig. 1, the system includes two parts, offline learning and online classification. Specially, offline learning mainly consists of five layers: heartbeat interval extraction, feature extraction, and feature analysis and model training.

4.2.1 Heartbeat interval extraction

We cut the whole night data into 5 min segments, then perform R wave detection and calculate the heartbeat interval, and correct it with R-wave error detection and miss detection algorithm.

Fig. 1 System framework



Algorithm 1 Improved heartbeat interval extraction algorithm (IHIEA)

Input: ECG Data sequence ECG={E1,E2,...,En},
Sliding window length winlen, window sliding step step, R wave peak point threshold PeakThrd, RR interval correction lower threshold lowDiffThrd, RR interval correction upper threshold highDiffThrd

Output:RR interval after 4Hz resampling ResampleRR

```
Peaks←Ø; %Save the position of the abscissa corresponding to the peak point of the R wave, initialize it to null
Diff←Ø   %save RR Intervel
UpdateDiff←Ø   %Save the error interval after the missed check RR period
ECGlen=n;
Pos=1:ECGlen; %The abscissa position corresponding to the data sequence
M ←floor(ECGlen/winlen); cnt=0;
ResampleRR←Ø;
while cnt<M do
  start←cnt*winlen;%Calculate the starting position of the cnt sliding window
  for   i←start+1 to start+winlen-2 do
    if ECG(i+1)>ECG(i)&&ECG(i+1)>ECG(i+2)&&ECG(i+1)>PeakThrd do
    %R wave peak threshold discrimination
      Peaks←Peaks∪{Pos(i+1)};
    end if
  end for
  start=start+step;
end while
% Detecting the R wave peak position of the remaining data sequence
start=M*winlen;
if start<ECGlen do
for   i←start+1 to ECGlen-2 do
if ECG(i+1)>ECG(i) &&ECG(i+1)>ECG(i+2)&&ECG(i+1)>PeakThrd do
   Peaks←Peaks∪{Pos(i+1)};
  end if
end for
end if
Peakslen=length(Peaks);
  for   i←2 to Peakslen do
        diflen=Peaks(i)-Peaks(i-1);%Calculate the difference between adjacent peak positions
  if diflen<DifThrd do %If the adjacent peak difference is less than the set threshold, it is wrong
     if Peaks(i)<Peaks(i-1) do%Remove the smaller peak from the set of peak locations
       remove(Peaks,Peaks(i));
     else   % Peaks(i)>Peaks(i-1)
          remove(Peaks,Peaks(i-1));
     end if
  end if
  end for
%Calculate the RR interval
Peakslen=length(Peaks);
for   i←2 to Peakslen do %Calculate the R wave peak difference and save
      diff←Peaks(i)-Peaks(i-1);
      Diff←Diff∪{diff};
end for
%RR interval miss detection algorithm
Difflen=length(Diff);
$\overline{Diff}$ =mean(Diff);
for   i←1toDifflen do
  if Diff(i)<lowDiffThrd* $\overline{Diff}$ ||Diff(i)>highDiffThrd* $\overline{Diff}$ do
     UpdateDiff ← UpdateDiff ∪{ $\overline{Diff}$ };
  else
         UpdateDiff ← UpdateDiff ∪{Diff(i)};
   end if
end for
ResampleRR ← Resample(UpdateDiff );
```

4.2.2 Feature extraction

To effectively extract and quantify the factors impacting hypertension recognition, we extract features from different perspectives including the linear domain features, nonlinear domain features, what's more, we use the Gaussian mixture model to model the distribution of the heartbeat interval sequence of each time window, and calculate the relevant parameters of the Gaussian mixture model as features.

4.2.3 Feature analysis

Considering that the linear domain and nonlinear domain features are from the same data source, and each of these features reflects the intrinsic properties of the cardiac autonomic nervous system from different aspects, it is believed that there may be some degree of correlation between these features, so the correlation between linear domain and nonlinear domain features is analyzed.

4.2.4 Algorithm selection

With these features extracted from heartbeat interval sequence, we use a variety of commonly used classification algorithms to train the model, and compare the performance of various classification algorithms. Finally, we choose random forest classification algorithm with better classification effect to train the hypertension recognition model.

5 Hypertensive patients recognition

In this section, we first obtain heartbeat interval sequence from ECG data, then extract features from different perspectives from heartbeat interval sequence to characterize different properties. Next, we use the Pearson correlation analysis method to quantitatively analyze the correlation between linear and nonlinear domain features and select features based on correlation between features and information gain. Finally, we train a hypertensive recognition model based on multi-dimensional features.

5.1 Heartbeat interval extraction

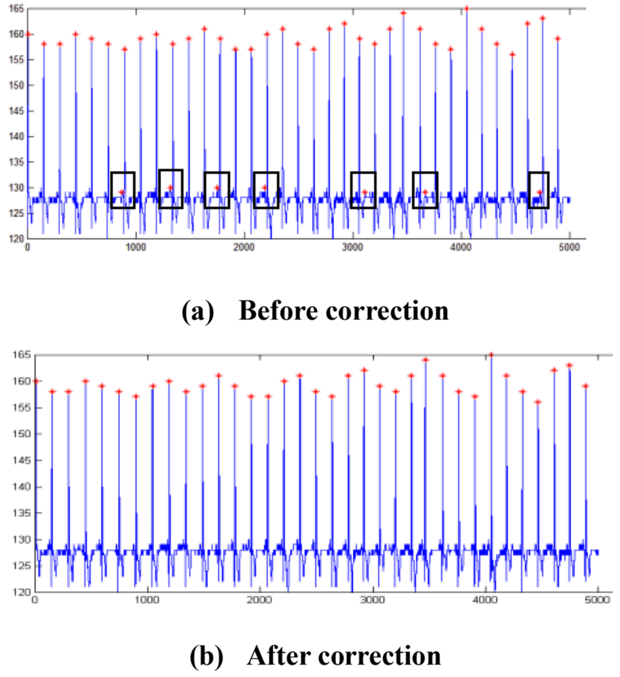
The heartbeat interval data is the basis for studying hypertension and other related diseases by using heart rate variability. In this paper, the heartbeat interval indicates the difference between the peak positions of adjacent R waves in the ECG, and the R wave peak is accurately detected. Thereafter, the time interval between adjacent R waves is the RR interval. Therefore, accurately detecting the position of the R wave peak is the basis for accurately calculating the heartbeat interval. We use the fixed-length sliding window method to detect the position of the R wave peak in this paper. When the peak detection method based on the sliding window is used for peak detection, the accuracy of the detection of the peak point has a certain relationship with the width of the window. When the window is large, the length of the covered data segment is long. When the peak detection is performed, the peak point is missed, and the peak point of some R waves cannot be recognized (ie, the R wave is missed); when the window is small, the length of the data segment is short, and the detection of the peak causes a misdetection of the peak point, and the point that is not part of the R wave is recognized as the peak point (ie, the R wave is misdetected). Therefore, how to effectively detect the peak value of R wave, reduce the occurrence of false detection and missed detection, and ensure the correct extraction of the heartbeat interval sequence is a problem worth studying.

In this paper, through many experiments, it is found that the window length is set to 200 and the moving step is set to 110, which can effectively reduce the occurrence of false detection and missed detection. At the same time, the experiment found that although the reasonable setting of the window length and the step length can effectively reduce the false detection and missed detection of the R wave peak, misdetection and missed detection still occur. Therefore, this paper proposes R wave Error detection and miss detection correction algorithms to further reduce the probability of occurrence of misdetection and missed detection based on the specific features of the ECG signal. as shown in Algorithm 1, where the error detection correction algorithm corresponds to lines 8 to 37, and the miss detection algorithm corresponds to lines 39 to 43. Figure 2a shows the R-wave diagram detected by using the sliding window. It can be seen that there are many false detections of R waves, and the points that are not R waves are erroneously detected as R waves. Figure 2b shows the R wave position using R wave misdetected correction algorithm. The correction algorithm performs a R-wave peak detection on the same piece of data, and it can be seen that the problem of false detection of the R wave is effectively solved. Figure 3 shows the R-wave miss detection. Unlike the error detection, the missed detection cannot be avoided by setting the threshold. Considering when the missed detection occurs, the interval between two adjacent R-waves becomes larger. Therefore, the missing test is corrected by the RR interval mean interpolation method.

5.2 Feature extraction

5.2.1 Time domain features

The time domain feature refers to the statistical analysis of the variation of the heartbeat interval over a period of



(a) Before correction

(b) After correction

Fig. 2 R wave error check

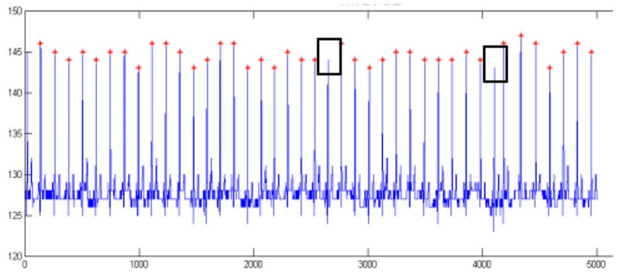


Fig. 3 R wave miss check

time, also known as statistical features. we extract eight features from heartbeat interval sequence as follows.

(1) Mean, indicating the average of the heartbeat interval sequence, reflecting the average level of the heartbeat interval sequence. The formula is as follows:

$$mean = \frac{1}{n}\sum_{i=1}^{n} RR_i \tag{1}$$

(2) The maximum value max represents the maximum value of the heartbeat interval sequence, and the calculation formula is:

$$MAX = \max\left(RR_1, RR_2, RR_3, \ldots, RR_n\right) \tag{2}$$

(3) The minimum value min indicates the minimum value of the heartbeat interval sequence, and the calculation formula is:

$$MIN = \min\left(RR_1, RR_2, RR_3, \ldots, RR_n\right) \tag{3}$$

(4) The standard deviation indicates the standard deviation of the heartbeat interval sequence, reflecting the degree of dispersion of the heartbeat interval sequence. The calculation formula is:

$$SD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(RR_i - \overline{RR}\right)^2} \tag{4}$$

(5) The root mean square of the difference, the square root of the mean squared difference of the lengths of all adjacent adjacent heartbeats, calculated as:

$$RMSSD = \sqrt{\frac{1}{n}\sum_{i=2}^{n}\left(RR_i - RR_{i-1}\right)^2} \tag{5}$$

(6) Difference standard deviation sdsd: indicates the standard deviation of all RR interval differences over a period of time. The calculation formula is as follows:

$$SDSD = \sqrt{\frac{1}{n}\sum_{i=2}^{n}\left[\left(RR_i - RR_{i-1}\right) - \overline{RR_i - RR_{i-1}}\right]^2} \tag{6}$$

(7) The number of neighboring RR intervals greater than 50 ms is nn50, the number of adjacent cycles above 50 ms interval.

$$nn50 = sum(RR \geq 50) \tag{7}$$

(8) The coefficient of variation cv, the degree of variation of the heartbeat interval sequence over a period of time. This feature is normalized by the mean value to standard deviation, which can offset the impact of individual differences. The calculation formula is:

$$\mathrm{CV} = \mathrm{SDNN}/\mathrm{mean} \tag{8}$$

5.2.2 Frequency domain features

The information obtained from the time domain analysis of the heartbeat interval sequence is finite and cannot fully reflect the attributes of the data sequence. Therefore, this section considers the features of the time series from the frequency domain. We use the fast Fourier transform method (Clifford and Tarassenko 2002) to convert the heartbeat interval sequence into a frequency signal and extract five features from the power spectrum density.

(1) High-frequency HF: (0.15 ~ 0.4 Hz) high-frequency band energy value, mainly related to the activity of

parasympathetic nerves, reflecting the rapid changes in heart rate.

(2) Low frequency LF: (0.04 ~ 0.15 Hz) The energy value of the low frequency band is mainly affected by sympathetic nerve and parasympathetic nerve, among which sympathetic nerve is dominant.

(3) Low frequency high frequency ratio LF/HF: reflects the balance of sympathetic and parasympathetic activity.

(4) Normalized high frequency $\mathrm{HF}_{\mathrm{norm}}$:

$$HF_{norm} = \frac{HF}{HF + LF} \tag{9}$$

(5) Normalized low frequency $\mathrm{LF}_{\mathrm{norm}}$:

$$LF_{norm} = \frac{LF}{LF + HF} \tag{10}$$

5.2.3 Nonlinear domain features

Poincare's scatter plot (France and Miroljub 2002) are often used to identify hidden patterns in time-series signals, and are one of the most common methods for detecting complex nonlinear behavior in heart rate variability. The Poincare scatter plot for hypertensive patients and healthy people is shown in Fig. 4. The scatter plot is based on the current heartbeat interval as the abscissa value and the next heartbeat interval as the ordinate value. Through observation, it is found that the scatter plot of healthy people is compact, 45-degree oval, mainly distributed in the middle of the ellipse, with less distribution at both ends, while the scatter plot of hypertensive patients is more scattered and has no regular shape. It is speculated that this reason may be that the physiological state of healthy people is stable, the range of heartbeat interval is small, and the expression in the scatter plot is relatively concentrated, while the hypertensive patients are sympathetically activated, and the ratio of sympathetic and parasympathetic nerves is unbalanced.

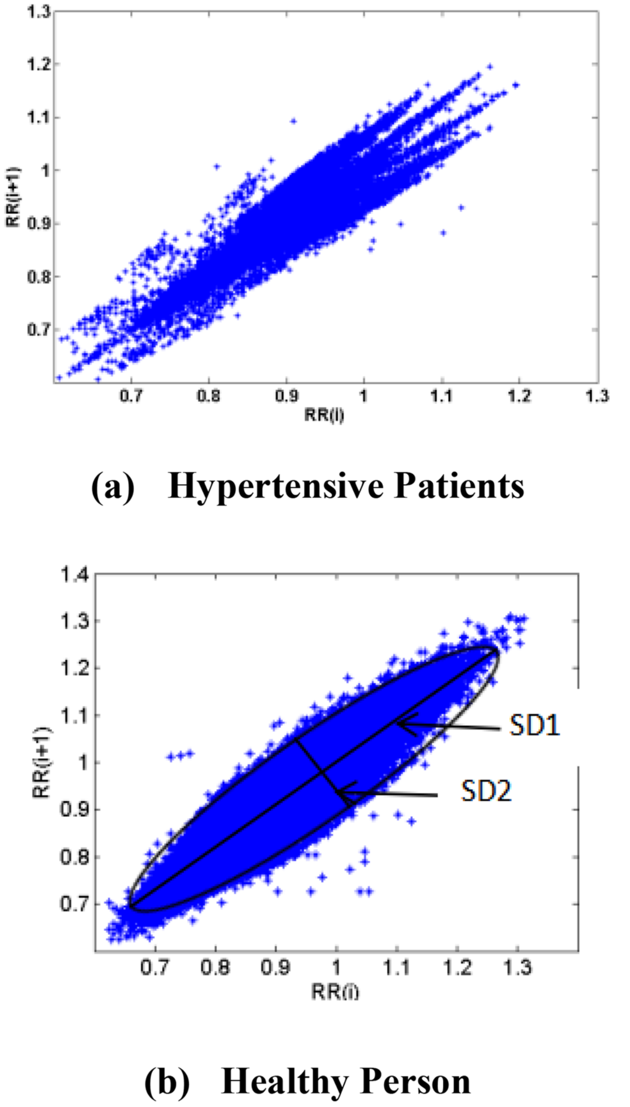
In order to quantitatively analyze the properties of the scatter plot, the ellipse is commonly used to fit the scatter plot, and then the standard deviation on the major and minor axes at the center of the scatter plot is determined. The center point is the X and Y axis mean heartbeat interval. The intersection of the lines, usually located on the 45-degree line of the coordinates, the 45-degree line representing the long axis of the scatter plot, and the vertical intersection of the long-axis at the center point is the short axis. As shown in Fig. 4b, the standard deviation measured by the long axis is called SD1, which reflects the difference between heartbeat and heartbeat. It is controlled by sympathetic nerves. The standard deviation measured by the short axis is called SD2. It reflects the difference in the RR interval over a long period of time and is controlled by the parasympathetic nerve. In addition, the relative size of the instantaneous and long-term RR interval differences is evaluated by SD1/SD2.

5.2.4 Distribution morphological features

In this section we analyze the properties of the histogram of the sequence of the center of each time window. Firstly, we calculate skewness and kurtosis (Xiaonan et al. 2018) of the histogram of heartbeat interval sequence. Skewness reflects the symmetry of the distribution pattern of heartbeat interval.

$$s = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(RR_i - \overline{RR}\right)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}\left(RR_i - \overline{RR}\right)^2\right]^2} \tag{11}$$

The kurtosis is a statistic that describes the degree of steepness in the distribution of the values of the variables.



(a) Hypertensive Patients

(b) Healthy Person

Fig. 4 The Poincare scatter plot for hypertensive patients and healthy people

When the data distribution is the same as the standard normal distribution, the kurtosis value is equal to 0.

$$k = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(RR_i - \overline{RR}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(RR_i - \overline{RR}\right)^2\right)^2} - 3 \tag{12}$$

In the previous section, it is observed that the distribution of heartbeat intervals in each time window exhibits a double peak, similar to the shape of the Gaussian Mixture Model (GMM) (Chen et al. 2016; Costa et al. 2012), and existing research indicates that people are within a certain period of time, the distribution of parameters such as heart rate and pulse rate is consistent with the Gaussian mixture model. Therefore, the Gaussian mixture model is used to model the distribution of heartbeat intervals in each time window.

Because finite hybrid models can be used to define any complex probability density function, it is widely used in many cases where statistical data is modeled. Statistically speaking, the premise of using the Gaussian mixture model is to assume that the sample data conforms to the independent and identical distribution, using a linear combination of normal distributions to approximate the unknown distribution of the data. The Gaussian mixture model is based on the following probability density hypothesis: Given a sequence of sample data, the model considers the probability of occurrence of each sample as a result of a mixture of several Gaussian models, which are sampled from the same probability density function while independent of each other. Usually a Gaussian mixture model is composed of K division Gaussian models, and the probability density function of the Gaussian mixture model is a linear superposition of K divisional Gaussian models. The probability density expression of the Gaussian mixture model as follows.

$$f(x) = w_1 * \frac{1}{\sqrt{2\pi\sigma_1}} \exp\left\{-(x-\mu_1)^2/2\sigma_1^2\right\} + w_2 * \frac{1}{\sqrt{2\pi\sigma_2}} \exp\left\{-(x-\mu_2)^2/2\sigma_2^2\right\} \tag{13}$$

where u_i and σ_i represents the mean and standard deviation of the i-th Gaussian distribution, w_i represents the weight of each Gaussian distribution in the GMM, and w_i satisfies the following constraints.

$$0 \le w_i \le 1, \sum_{i=1}^{2} w_i = 1, i = 1, 2 \tag{14}$$

The probability density expression for a single Gaussian distribution as follows.

$$f\left(x_i|\theta_i\right) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{\left(x_i - u_i\right)^2}{2\sigma_i^2}\right\} \tag{15}$$

In order to obtain the optimal solution of unknown parameters, the traditional method is solved by the method of maximum likelihood estimation. The specific solution step is to find the joint probability density function of the sample first, because each sample is independent and identically distributed. Therefore, these samples are the joint probability density is the product of the probability density of a single sample and is expressed as:

$$L(\theta) = \prod_{i=1}^{K} f\left(x_i|\theta_i\right) \tag{16}$$

Next, taking the logarithm of the joint probability density function,

$$LL\left(\theta_i\right) = \log^{\prod_{i=1}^{K} f\left(x_i|\theta_i\right)} = \sum_{i=1}^{m} \log^{f\left(x_i|\theta_i\right)} \tag{17}$$

Then use the maximum likelihood estimate to determine the optimal value of the unknown parameter $\theta_i = \left\{\mu_i, \sigma_i\right\}$,

$$(\mu, \sigma)_{MLE} = \arg\max_{\mu,\sigma} LL(\mu, \sigma) = \arg\max_{\mu,\sigma} \sum_{i=1}^{m} \log^{f\left(x_i;\mu,\sigma\right)} \tag{18}$$

Since the above formula contains the logarithm of the sum, it is difficult to maximize the log-likelihood function. In order to solve the above problem, Demptater proposed the expectation maximization (EM) algorithm (Zhihua 2016) in 1977, A commonly used algorithm in the field of machine learning domain, which can effectively solve the optimization problem of hidden variables in Gaussian mixture models.

In this study, the specific process of the algorithm includes the following four steps.

(1) Given the number K of individual Gaussian distributions in the mixed Gaussian model, assigning initial values to the parameters w_i, u_i, σ_i of each Gaussian distribution;
(2) E-step: Calculate the posterior probability of the hidden variable w according to the initial value of the parameter or the model parameter of the last iteration, which is essentially the expectation of the invisible variable as the current estimated value of the hidden variable;
(3) M-step: Calculate the value of the hidden variable as an input by E-step, and maximize the likelihood function to calculate the values of the new parameters u, σ and w;
(4) Iterating 2) and 3) until the value of u, σ remains unchanged, at which point the maximum value of the parameters u, σ can be obtained.

After calculating the parameters of the Gaussian mixture model, this paper selects the mean, standard deviation and the corresponding probability density of the two Gaussian distributions as the features for the recognition of hypertension.

5.3 Feature correlation analysis of linear domain and nonlinear domain

Considering that the linear domain and nonlinear domain features extracted from this paper are from the same data source, and these features reflect the intrinsic properties of the cardiac autonomic nervous system from different aspects. therefore, these features are likely to be related. We use the Pearson correlation coefficient to analyse the correlations between time domain and frequency domain features Quantitatively. The Pearson Correlation Coefficient (PCC) is a quantitative measure of the degree of correlation between features (Wang et al. 2016). For two eigenvectors X and Y, $X=(X_1, X_2, \ldots, X_n)$, $F2=(Y_1, Y_2, \ldots, Y_n)$, the correlation coefficient PCC of these two features is calculated as:

$$PCC = \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(Y_i-\overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\left(Y_i-\overline{Y}\right)^2}} \tag{19}$$

The method for calculating the feature correlation coefficient of this paper as follows:

Step 1: For each person, we cut the whole night data into 5 min time window, and extract the linear domain and nonlinear domain features from the time window, and form the feature column vector F_i, ie $F_i=\{F_{i,1};F_{i,2};\ldots;F_{i,j}\ldots F_{i,96}\}$, get the feature matrix FM_k.

$$FM_k = \begin{bmatrix} F_{1,1} & F_{1,2} & . & . & F_{1,16} \\ F_{2,1} & F_{2,2} & . & . & F_{2,16} \\ . & . & & & . \\ . & . & & & . \\ F_{96,1} & F_{96,2} & . & . & F_{96,16} \end{bmatrix} \tag{21}$$

Step2: Calculate the Pearson correlation coefficient (PCC) for the two-two features in FM_k, and obtain the correlation coefficient matrix CM_k.

Step3: Calculate the correlation coefficient matrix in the same way for everyone, and average each coefficient in the correlation coefficient matrix of the same person as the correlation coefficient between the humanoid features.

$$CM_k = \begin{bmatrix} r_{1,1} & r_{2,1} & r_{3,1} & . & r_{16,1} \\ r_{1,2} & r_{2,2} & . & . & r_{16,2} \\ r_{1,3} & . & . & . & . \\ . & . & . & . & . \\ r_{1,16} & r_{2,16} & r_{3,16} & . & r_{16,16} \end{bmatrix} \tag{22}$$

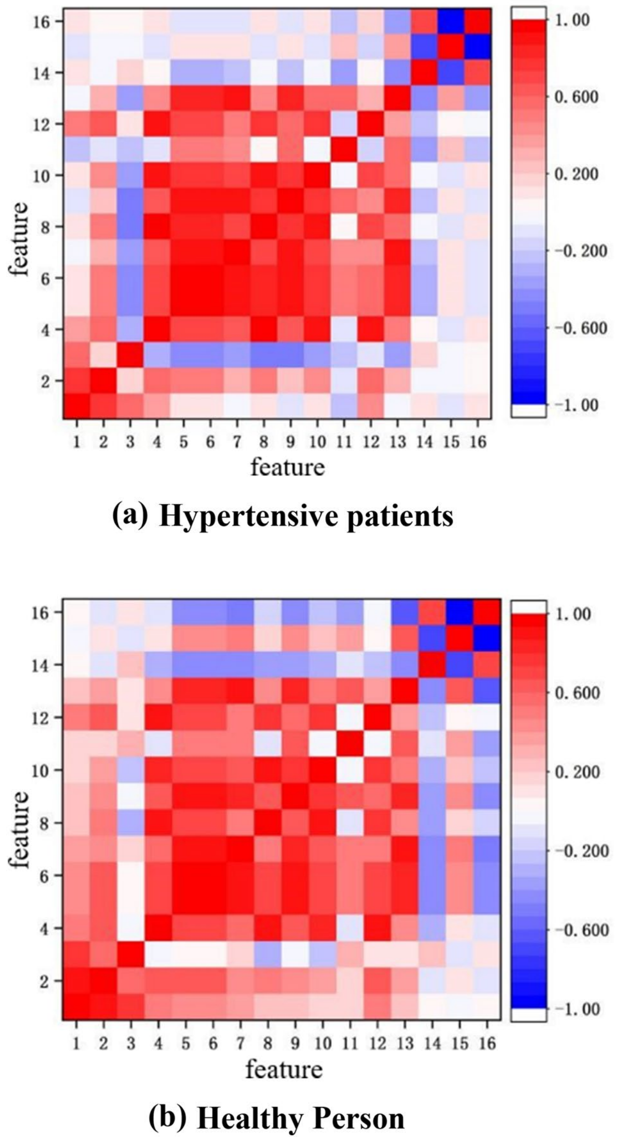
The feature correlation algorithm proposed in this paper is shown in Algorithm 2.

Algorithm 2 Feature correlation algorithm

```
Input: Features of 96 time windows for each type of person, number of persons N
Output: Correlation coefficient matrix AVG for each type of person
SUM ← Ø;AVG ← Ø;
for k=1:N
  $FM_k$ ← Ø;
  for i=1:16
    Fi ← Ø;
      for j=1:96
        $F_i$ ← $F_i$□ {$F_{i,j}$};
        $FM_k$ ← $FM_k$□FiT;
      end for
  end for
   for i=1:16
    for j=1:16
       $CM_{i,j}$ ← Correlate(FiT,FjT);
%Calculate the correlation coefficient of every two column vectors
       $CM_k$ ← $CM_k$ □{$CM_{i,j}$};
     end for
   end for
  SUM←SUM+$CM_k$;
  end for
AVG←SUM/N; %Average of the matrix of correlation coefficients for each type of person
```

We use the above method to calculate the correlation coefficient matrix between the features of hypertensive patients and healthy person, and the calculated correlation coefficient tables of various types of people are shown in Fig. 5. It can be seen that there is a strong correlation between some features, and there are some differences in the correlation strength between the two types of human features.

As can be seen from the above table, for the two types of people, the features correlated with the low frequency LF are the standard deviation SD, the coefficient of variation CV and the short axis standard deviation SD2 of the scatter plot, and there is also a strong correlation between the features related to the low frequency. The reason for the strong correlation between these features is that the low frequency is affected by the sympathetic nerve and the parasympathetic nerve. The sympathetic nerve is dominant, while the standard deviation and coefficient of variation are important to measure the slow change component of heart rate variability. SD2 reflects the long-term variability of heartbeat interval. These features are also affected by sympathetic nerves (Schroeder et al. 2003), and when the function of the sympathetic nerves changes, they change synchronously. In addition, the correlation between low-frequency and other features of hypertensive patients is higher than that of healthy people. The reason for this is related to the excessive activation of sympathetic nerve in hypertensive patients.



(a) Hypertensive patients

(b) Healthy Person

Fig. 5 Heatmap of correlation coefficient

The features strongly related to high frequency HF are the difference root mean square rmssd, the difference standard deviation sdsd, the number of the heartbeat interval greater than 50 ms, and the long axis standard deviation sd1. There is also a strong correlation between these features, indicating that the values of these features is increased or decreased synchronously. The reason why these features are strongly correlated with each other is that the high frequency range is 0.15–0.4 Hz, which is very close to the frequency of respiration. Breathing causes high frequency periodic fluctuation of the vagus nerve, which changes the length of the hop-by-hop heartbeat interval. While rmssd, sdsd, nn50, and sd1 respond to short-term changes in the heartbeat interval, these features are commonly affected by the vagus nerve. In addition, the overall correlation between high frequency and other features of hypertensive patients is weaker than that of healthy people, which may be related to the inhibition of parasympathetic function in hypertensive patients.

The low-frequency high-frequency ratio LHratio reflects the dynamic equilibrium state of the sympathetic nerve and the vagus nerve, and has a strong correlation with sd12 in the scatter plot. This is because sd12 is the ratio of sd1 and sd2, and also reflects the equilibrium state of sympathetic and parasympathetic nerves. The correlation between the normalized high frequency and low frequency features is very strong ($r = -1.00$, $p < 0.05$), but the correlation with other features is very weak.

When there is a strong correlation between features, the information represented by these features is redundant, and feature selection can be made based on the correlation strength of the features. Here, the features are first aggregated based on the correlation strength between the features, and the features can be aggregated into seven groups, as shown in Table 1. For each set of features, one of the features needs to be selected instead of other features of the group to achieve feature selection. In this paper, Information Gain (IG) (Kullback et al. 1959; Entropy 2002) is used to measure the distinguishing ability of each feature for two types of people. The information gain for each feature is then replaced with the other features of the set with the most information gain in each group.

As shown in the above table, the features can be aggregated into seven categories according to the correlation

Table 1 Feature grouping and information gain of each feature

GroupID	Features	InfoGain	GroupID	Features	InfoGain
1	LF	0.00855	3	**lhratio**	**0.10567**
	sd	**0.01037**		sd1/sd2	0.05595
	cv	0.00783	4	$\mathrm{HF}_{\mathrm{norm}}$	0.10567
	Sd2	0.02426		$\mathbf{LF}_{\mathbf{norm}}$	**0.10569**
2	**HF**	**0.0439**	5	**mean**	**0.04609**
	rmssd	0.0402	6	**max**	**0.04569**
	sdsd	0.0402	7	**min**	**0.36945**
	nn50	0.04082			
	Sd1	0.03716			

strength between the features. The feature with the largest information gain in each class represents the other features of the group, and finally the features included in the subset of the time domain and frequency domain features are selected. There are standard deviation, low frequency high frequency ratio, normalized low frequency, high frequency, average, maximum and minimum, as shown in bold.

5.4 Hypertensive patients recognition

5.4.1 Feature process

5.4.1.1 Feature merge When extracting the heartbeat interval, we first divide the 8-h ECG data into a 5-min time window, and then extract three types of features from each time window. For each person, each feature is a vector of 96 values, where multiple values for each feature need to be merged into one value. The problem here is for the feature vector, which method is used to fuse the feature vector into a feature value to reflect as much as possible the properties of all the values in the data sequence, and the final result is more representative.

$$Feature = \left\{F_1, F_2, \dots, F_{96}\right\} \xrightarrow{?} value \tag{22}$$

In order to fuse multiple feature vectors into one special diagnosis vector, Boureau et al. (2010), Ni et al. (2017) and others have solved the problem by pooling. The pooling method is commonly used in convolutional neural networks. The convolutional neural network (Kiranyaz et al. 2016; Liu et al. 2015) includes an input layer, a convolution layer, a sampling layer, a connection layer, and an output layer. The sampling layer is also called a pooling layer. The downsampling is based on the principle of local correlation, which can effectively reduce the amount of data while retaining useful information (Romanuke Vadim 2017). Commonly used pooling methods include finding the maximum, minimum, average, and sum of squares of all values. The method of calculating the mean and sum of squares and roots can consider the information of all eigenvalues, the method of maximum and minimum. An extreme value in the data sequence is substituted for the data sequence. Since the maximum and minimum values reflect the extremes of the data sequence and do not fully reflect the overall properties of the data sequence, the average takes into account all the elements in the data sequence and can reflect the average level of the data sequence, but ignores the data sequence. The details, squares, and roots take into account all the element values in the data sequence, and reflect the intrinsic properties of the data sequence compared to the averaging method. Therefore, when we aggregate multiple values of a single feature into a single value, we use the method of square sum rooting, as shown in the following formula.

$$value = \sqrt{F_1^2 + F_2^2 + \cdots + F_{96}^2} \tag{23}$$

5.4.1.2 Feature normalization We find that the range of values between the features varies greatly, and the values of some features differ by several orders of magnitude. Therefore, before using the feature training model of this paper, the features are normalized. Normalization refers to a linear feature transformation method that scales the value of a feature to a specific range, but does not change the distribution of the feature values. In this paper, the min–max normalization method (Patro and Sahu 2015; Mustaffa and Yusof 2011) is used to transform the numerical range of each feature into the (0, 1) interval, and then the model is trained with the normalized feature. Given heartbeat interval sequence $RR = \left\{RR_1, RR_2, RR_3, \cdots, RR_n\right\}$, The formula for calculating the transformed eigenvalue by the min–max method is as follows.

$$RR_i' = \frac{RR_i - RR_{\min}}{RR_{\max} - RR_{\min}}, i = 1, 2, \dots, n \tag{24}$$

In the above formula, RRmax and RRmin respectively represent the maximum and minimum values of the heartbeat interval sequence.

5.4.1.3 Construction of hypertension recognition model Given the feature set of two types of people, the existing machine learning method is used to train the model, and then when there is unknown type of user data, the trained model can be used to output the prediction result of the user category. The extracted heart rate variability features and corresponding categories are composed of training samples, and the sets of all the training samples of the two types of people constitute a training set. In the section of constructing the hypertension recognition model, the training set is used as the input of the classification algorithm.

The classification algorithms commonly used in machine learning include decision trees, random forests, Bayesian networks and multi-layer perceptrons. These papers use these algorithms to train models and then evaluate the performance of these models. We extracted a total of 13 linear domain features, 3 nonlinear domain features, and 8 distributed morphological features. The feature subsets obtained in the correlation analysis section include 7 features. In this paper, these features are combined into different feature set training models to obtain different recognition models. The performance of the model is evaluated from different angles, and the model with the best recognition effect is used as the hypertension recognition model.

6 Performance evaluation

6.1 Experimental setup

6.1.1 Evaluation metrics

For features evaluation, we measure the effectiveness of the extracted features using following metric:

The goal of this paper is to identify patients with hypertension from healthy people. We take hypertension as a positive example. So we use Presicion, Recall and AUC (Wang et al. 2016) to evaluate the performance of the hypertension recognition model. AUC is an independent indicator, the larger the value, the better the recognition of the trained model. Indicates how many of the identified hypertensive patients do have high blood pressure, and the recall rate indicates how many hypertensive patients are correctly identified. The accuracy rate needs to be used in conjunction with the recall rate, when both indicators have large values at the same time. It shows that the system recognition effect is good.

$$\mathrm{Pr}\,ecision = \frac{TP}{TP + FP} \tag{25}$$

$$\mathrm{Rec}all = \frac{TP}{TP + FN} \tag{26}$$

In the above two formulas, TP indicates that it is originally a hypertensive patient, and is actually divided into hypertensive patients. FP indicates that it is originally a healthy person and is misclassified as a hypertensive patient. FN indicates that it is originally a hypertensive patient and is misclassified as a healthy person.

6.1.2 Verification method

In order to fully verify the performance of the proposed hypertension method, this paper used 10 tenfold cross-validation experiments. Specifically, for each tenfold cross-validation, the data set is randomly divided into 10 equal parts, each data is rotated as a test set and the remaining nine data are used as a training set.

6.1.3 Baseline algorithms

For hypertension recognition, we use the following methods as the baselines:

(1) In the work of Ni et al. (2019), the data of the whole night is first divided into 1/2, 1/4, 1/8, 1/16, 1/32 of the original length, which is equivalent to dividing the data into 6 layers, each layer contains several data segments, and the lengths of the data segments are equal. Then, for each piece of data of each layer, 20 features of time domain, frequency domain and entropy are extracted, and then the sum of squares for each feature is used to find the root. A plurality of feature vectors in each layer are fused into a single feature vector, and between the layers, a plurality of feature vectors are fused into a single feature vector by averaging. In the feature selection part, the information gain is obtained for the feature, and then the first seven feature training recognition models are selected;

(2) In the work of M. G. Poddar et al., a 5-min ECG record was used as the data source, and the features of the time domain, the frequency domain and the nonlinear domain were extracted respectively, and the support vector machine was used for classification. In this experiment, in order to make the experimental results comparable, the method described in this paper and the two control methods use the same experimental data.

6.2 Experimental results

6.2.1 Difference analysis of heart rate variability features

In order to analyze whether the features extracted in this paper have significant differences between the two types of people, the time domain frequency domain subsets and distribution morphological features obtained by two kinds of human correlation analysis are separately calculated. The statistical results of the features are expressed by the mean ± standard deviation in Table 2.

From above Table we can observe:

(1) Observing the eight features of the time domain, we found that in addition to the minimum and coefficient of variation, other time domain features of hypertensive patients are higher than those of healthy people. The three features of scatter plots can be found in patients

Table 2 Heart rate variability feature statistics

Feature	Hypertension	Heathy	p value
mean	0.998±0.099	0.871±0.084	2.77207E−08
sd	0.088±0.018	0.087±0.023	0.00729
max	1.35±0.089	1.158±0.089	1.08956E−09
min	0.599±0.049	0.701±0.051	2.01966E−14
rmssd	0.025±0.011	0.0191±0.007	0.00729
sdsd	0.025±0.011	0.0191±0.007	4.31874E−09
nn50	10,193.159±9461.904	6351.891±5675.355	3.82187E−08
cv	0.087±0.019	0.101±0.017	3.82153E−08
LF	0.00093±0.00074	0.00087±0.00054	0.00284
HF	0.00005±0.00002	0.000038±0.00001	5.45688E−05
lhratio	527.722±112.242	558.568±163.608	6.08009E−06
hfnorm	0.00198±0.00046	0.00157±0.00035	1.27906E−06
lfnorm	0.998±0.00046	0.998±0.00035	0.00214
sd1	9203.798±2956.413	10,209.345±2577.855	3.96339E−08
sd2	29,528.719±5580.647	36,922.358±4814.064	5.6803E−05
sd12	0.277±0.064	0.310±0.067	1.23698E−09
mainU	1.004±0.106	0.877±0.088	3.68111E−10
subU	0.969±0.112	0.822±0.119	4.77362E−07
mainstd	0.078±0.019	0.075±0.028	4.76056E−05
substd	0.057±0.038	0.042±0.029	3.38771E−06
mainweight	0.764±0.130	0.798±0.132	7.32134E−06
subweight	0.236±0.131	0.202±0.132	9.92831E−06
skew	0.073±0.571	0.067±0.346	4.26276E−05

with hypertension. All three features are lower than healthy people. In addition, the p-values of these 11 features of hypertensive patients and healthy people are less than 0.05, and the P value of some features is even less than 2E−5. Such a small p-value indicates that the time domain features of hypertensive patients and healthy people are derived from the probability of the same distribution is small.

(2) Observing the five features of measuring the frequency domain information of heartbeat interval, we found that except for the low frequency high frequency ratio and the normalized low frequency, other features of hypertensive patients are higher than those of healthy people. Similarly, the T-test results of these two types of people are less than 0.05, which indicates that the frequency domain features of the two types of people are from the same distribution with a low probability.

(3) Observing the distribution morphological features, except for the weight and kurtosis coefficient of the main Gaussian distribution, other features of hypertensive patients are higher than those of healthy people. The P value of these characteristic T tests is much smaller than 2E−5, such a small p value indicates the time domain features of hypertensive patients and healthy people are small from the same distribution.

6.2.2 Comparison of different algorithms

We want to compare the effectiveness of different algorithms in hypertension recognition. We use all the extracted features, train the models with four classification algorithms: random forest (RF), decision tree (DT), Bayesian network (BN), and multi-layer perceptron (MP), and compare the accuracy, recall and AUC of these models.

Figure 6 shows the Precision, Recall and AUC of RF, DT, bN and MP. We observe that the classification precision and

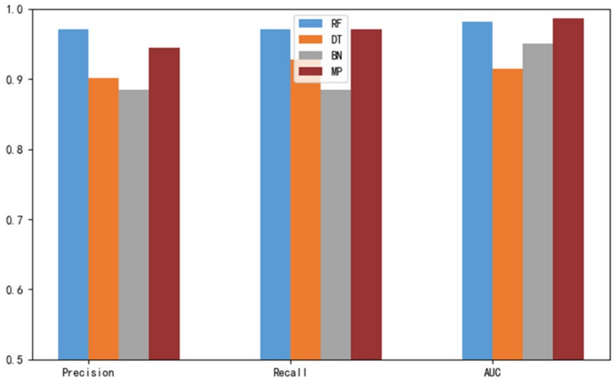


Fig. 6 Comparison of different algorithms

recall of the four types of algorithms can reach above 0.85 When using all feature training models. Among them, the RF has the highest precision and recall rate and is used as the default classification algorithm to construct a classification model. This shows that when the features proposed in this paper are used to identify hypertensive patients, it can ensure high recall rate while ensuring high detection rate.

6.2.3 Performance comparison before and after feature selection

In order to compare the performance of data sets before and after feature selection, we compare the performance of different feature sets here. the feature sets here include four categories: (1) Linear and nonlinear domain feature sets (TF); (2) Correlation analysis obtained feature subsets (TFsubset); (3) Linear domain, nonlinear domain and distribution morphological feature set (TF+RRHis); (4) a collection of feature subsets and distribution patterns (TFsubset+RRHis).

Figure 7 shows the Precision, Recall and AUC of four types of feature sets. It can be seen that compared with the simultaneous use of the all features, the classification precision is not significantly reduced by using the subset training model. This is because some features of the linear domain and the nonlinear domain are strongly correlated, and the information of the strongly correlated features is redundant, and the removal of some features does not lead to information loss. When three types of features are used at the same time, the attributes of the heartbeat interval sequence can be reflected from different angles, so the recognition result is the best.

6.2.4 Comparison of different type of feature sets

What's more, we extract three types of feature sets from heartbeat interval sequence, here we compare the performance of different types of feature sets. the feature sets include three categories: (1) Linear and nonlinear domain feature sets (TF); (2) Linear and nonlinear domain feature subsets (TFsubset); (3) Distribution morphological features (RRHis).

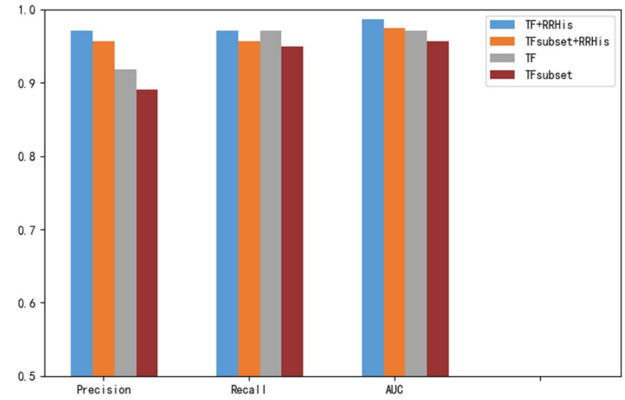


Fig. 7 Comparison of before and after feature selection

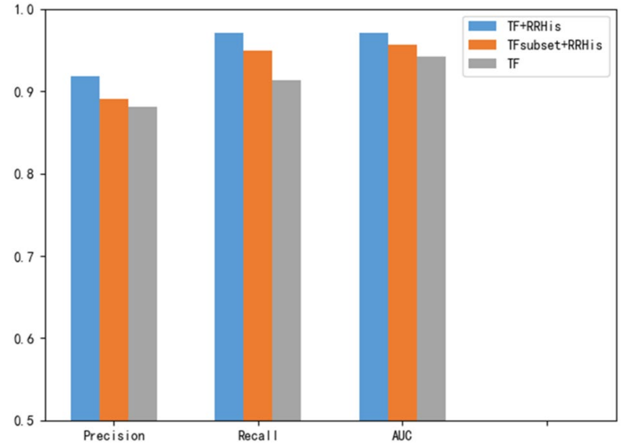


Fig. 8 Recognition result

As can be seen from the above Fig. 8, the accuracy, recall, and AUC of the hypertension recognition model trained using linear and nonlinear domain characteristics are 91.8%, 97.1%, and 97.1%, respectively, and the performance of the hypertension recognition model trained using the distribution pattern features. It is slightly worse than the other two types of feature sets. The reason for the analysis is that the distribution morphological feature is a coarse-grained description of the distribution pattern of the heartbeat interval sequence, while the linear and non-linear domain features reflect the changing trend of the corresponding heartbeat interval sequence, which is a fine-grained feature. Therefore, the latter classification Better results.

6.2.5 Comparison of principal component analysis

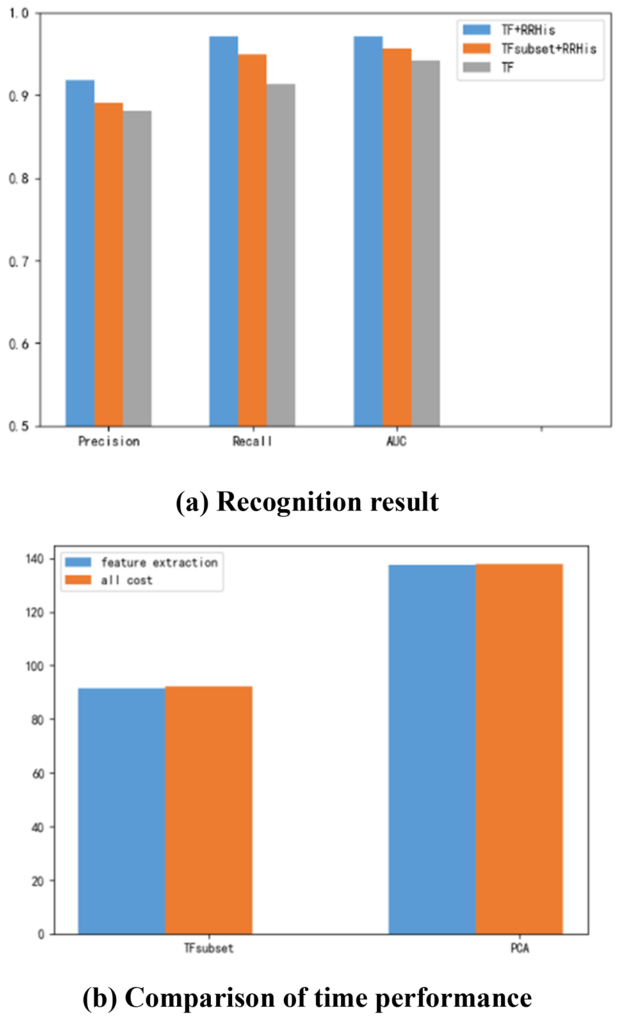
In order to verify the validity of the feature selection method proposed in this paper, it is compared with the principal component analysis method. We use all linear domain and nonlinear domain features as the input of the principal component analysis method to obtain the principal component. In order to keep the number of features obtained from the correlation analysis in this paper, the first seven principal component training models are used here. The proportion of each of the main components is shown in Table 3. It can be seen that the cumulative weight of the first nine principal components reaches 95.7553% (Table 3).

We denote the seven principal component sets obtained from principal component analysis as PCA, train the models with two types of feature sets PCA and TFsubset, and compare the performance of the two models.

Table 3 Principal components and weight

Principal component	Weight (%)	Principal component	Weight (%)
pc1	49.2429	pc5	3.9577
pc2	17.8195	pc6	2.5120
pc3	10.9921	pc7	2.1393
pc4	9.0918	Cumulative sum	95.7553

Figure 9 shows the Precision, Recall and AUC of TFsubset and PCA. It can be seen that the accuracy, recall rate and AUC of the model trained using TFsubset are slightly higher than PCA, and the features selected in this paper have clear physiological significance, and are highly explanatory, which is helpful to understand which features are more important. and the physiological significance expressed by the principal component obtained by principal component analysis is not clear. Comparing the time performance of these two methods, the time required to calculate the features of 276 individuals using PCA is 45.92 s longer than the time required to calculate the subset of time domain frequency domain features, while the decision tree is used to train the models with these two feature sets. The time required is less than 1 s. In summary, the feature subset of this paper can reduce the time performance while achieving higher accuracy.



(a) Recognition result

(b) Comparison of time performance

Fig. 9 Performance comparison in model and time

6.2.6 Comparison of different baseline methods

In order to confirm the validity of the method proposed in this paper, we compare it with the methods of Ni et al. (2017) and Poddar et al. (2014).

From the above Table 4, we can observe that the precision and recall rate of the method were 7.1% and 3.4% higher than the first baseline, and 18.4% and 9.2% higher than the second baseline.

We compared the ROC curves of the three types of methods for hypertension recognition. As can be seen from the Fig. 10, the method described herein achieves a relatively low false alarm rate while ensuring a relatively high recognition rate relative to the two control methods. Specifically, the best case of this method can guarantee a recognition rate of 98.6%, while the false positive rate is only 4.35%, while the method of NI et al. can achieve a recognition rate of 94.2% in the best case. At the same time, the false positive rate was 12.3%. In the best case of MG Poddar et al., the recognition rate was 87.7% and the false positive rate was 23.9%.

Table 4 Comparison of recognition result of different methods

	Precision	Recall	AUC
Ni 2017	0.900	0.935	0.866
Poddar 2014	0.787	0.877	0.769
Our	0.971	0.969	0.987

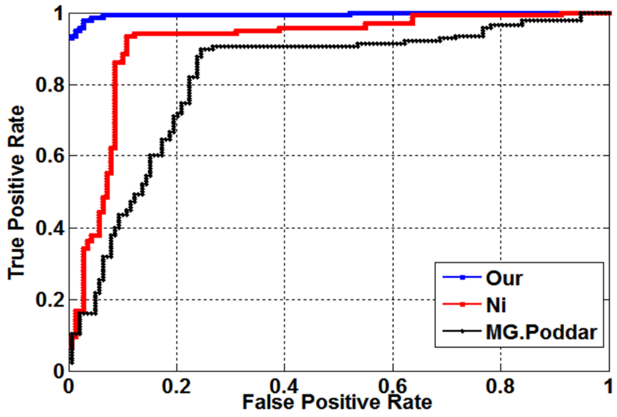


Fig. 10 ROC Curve of Hypertension recognition

6.2.7 Analysis of time complexity

Time complexity is an effective indicator to measure the performance of the proposed method. In order to analyze the time performance of the proposed hypertension method on different scale data sets, this part we compare the data sets of different scales in the heartbeat interval extraction. The time consumed by feature extraction and the total CPU consumption time are shown in Fig. 11. Compared with heartbeat interval extraction and feature extraction, the time required to train the model is much lower than the two processes. Therefore, the time performance when training the model is not considered here. In the experiment, we tested the time performance and total time performance of extracting heartbeat intervals, extracting features using different numbers of ECG signals.

We can observe that whether the time to extract the heartbeat interval or features, or the total CPU consumption time of these two parts, it has a linear relationship with the data size, and does not increase sharply with the increase of the data set size.

7 Conclusion

In this paper, we focus on the problem of recognizing hypertensive patients from healthy person leveraging ECG data. Specifically, we first extract heartbeat interval sequence from ECG data and propose an improved heartbeat interval extraction algorithm to solve the problem of false detection and missed detection of R waves. Then we extract features from different perspectives including the linear domain, nonlinear domain and Distribution morphological features from heartbeat interval sequence. We propose a method of modeling the distribution pattern of heartbeat interval in each time window based on Gaussian mixture model. Considering the correlation between linear domain and nonlinear domain features, we use Pearson correlation analysis method to analyze the correlation strength quantitatively, and make feature selection based on correlation strength and information gain. Finally, Based on the features of multi-dimensional heart rate variability, an early recognition model of hypertension based on random forest classification algorithm is constructed. Experimental results demonstrate the effectiveness of our approach.

In the future work, we will consider using data from each person for a period of time before and after the onset of illness, and based on the fluctuation pattern of heart rate variability in each person during this period, establish a predictive model. Besides, we will examine the correlation between heart rate variability features and two types of people to select targeted features.

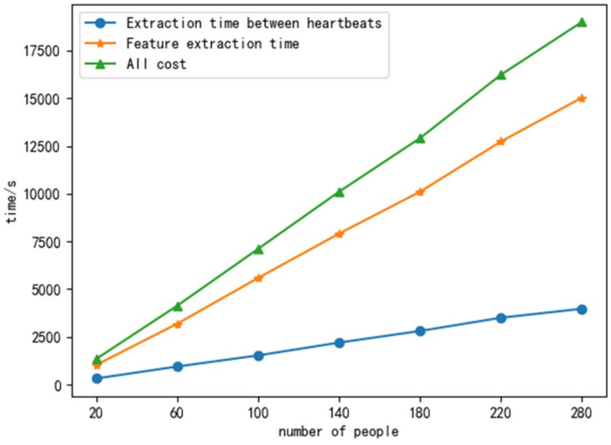


Fig. 11 Time performance curve

Acknowledgements We thank the reviewers for the valuable comments and spending time towards the improvement of the paper. Meanwhile, we should also thank the subjects to help us conduct the presented experiments. This work is supported by State Key Program of National Natural Science of China (No. 61332013).

References

Boureau Y, Ponce J, LeCun Y (2010) A theoretical analysis of feature pooling in visual recognition. ICML

Carek A, Holz C (2018) Naptics: Convenient and continuous blood pressure monitoring during Sleep. Proc ACM Interact Mob Wearable Ubiquitous Technol 2(3):96. https://doi.org/10.1145/3264906

Carek AM, Conant J, Joshi A, Kang H, Inan OT (2017) SeismoWatch: wearable cufess blood pressure monitoring using pulse transit time. Proc ACM Interact Mob Wearable Ubiquitous Technol 1(3):1–16. https://doi.org/10.1145/3130905 (**Article 40**)

Chen Y, Chen W, Kitamura K, Nemoto TT, Guiyun (2016) Long-term measurement of maternal pulse rate dynamics using a home-based sleep monitoring system. J Sens. https://doi.org/10.1155/2016/5730142

Clifford GD, Tarassenko L (2002) Signal processing methods for heart rate variability. University of Oxford, Oxford

Coccagna G, Lugaresi E (1978) Arterial blood gases and pulmonary and systemic arterial pressure during sleep in chronic obstructive pulmonary disease. Sleep 1(2):117–124

Costa T, Boccignone G, Ferraro MB, Jérémie (2012) Gaussian mixture model of heart rate variability. PLoS One 7(5):e37731

Couderc JP, Xiaojuan X, Zareba W, Moss AJ (2005) Assessment of the stability of the individual-based correction of QT interval for heart rate. Ann Noninvasive Electrocardiol 10(1):25–34

DataBase (2015) https://physionet.org/physiobank/database/shareedb/

Davies RJO, Jenkins NE, Stradling JR (1994) Effect of measuring ambulatory blood pressure on sleep and on blood pressure during sleep. BMJ 308(6932):820–823

Fletcher EC, Levin DC (1984) Cardiopulmonary hemodynamics during sleep in subjects with chronic obstructive pulmonary disease: the effect of short-and long-term oxygen. Chest 85:6–14

France S, Miroljub J (2002) Heart rate variability—a shape analysis of Lorenz plots. Cell Mol Biol Lett 7(1):159–161

Information_Entropy (2002) https://en.wikipedia.org/wiki/Entropy_(information_theory)

Kiranyaz S, Ince T, Gabbouj M (2016) Real-time patient-specific ECG classification by 1-D convolutional Neural Networks. IEEE Trans Biomed Eng 63(3):664–675

Kullback S (1959) Information theory and statistics[M]. Wiley, Hoboken

Li W, Gu H, Teo KK, Bo J, Wang Y, Yang J, Wang X, Zhang H, Sun Y, Jia X et al (2016) Hypertension prevalence, awareness, treatment, and control in 115 rural and urban communities involving 47000 people from China. J Hypertens 34(1):39–46

Liu T, Fang S, Zhao Y et al (2015) Implementation of training convolutional neural networks. arXiv preprint arXiv:1506.01195

Mustaffa Z, Yusof Y (2011) A comparison of normalization techniques in predicting. 2010 international conference on business and economics research, vol 1. IACSIT Press, Kuala Lumpur

NCD Risk Factor Collaboration (NCD-RisC) (2016) Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19·1 million participants. Lancet 2017(389):37–55. https://doi.org/10.1016/S01406736(16)31919-5

Ni H, Cho S, Mankoff J, YangDey JAK (2017) Automated recognition of hypertension through overnight continuous HRV monitoring. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-017-0471-y

Ni H, Wang Y, Xu G, Shao Z, Zhang W, Zhou X (2019) Multiscale fine-grained heart rate variability analysis for recognizing the severity of hypertension. Comput Math Methods Med 2019:1–9. https://doi.org/10.1155/2019/4936179 (**Article ID 4936179**)

Patro SGK, Sahu KK (2015) Normalization: a preprocessing stage. arXiv:1503.06462

Poddar MG, Kumar V, Sharma YP (2014) Heart rate variability based classifcation of normal and hypertension cases by linear-nonlinear method. Def Sci J 64(6):542–548. https://doi.org/10.14429/dsj.64.7867

Romanuke Vadim V (2017) Appropriate number of standard 2×2 max pooling layers and their allocation in convolutional neural networks for diverse and heterogeneous datasets. Inf Technol Manag Sci 20(1):12–19

Rui W, Weichen W, Alex D, Jeremy H, William K, Todd H, Andrew C (2018) Tracking depression dynamics in college students using mobile phone and wearable sensing. Proc ACM Interact Mob Wearable Ubiquitous Technol 2(1):1–26

Schroeder B, Emily LE, Duanping CJ, Lloyd PW, Ronald EW, Gregory HW, Gerardo R (2003) Hypertension, blood pressure, and heart rate variability the atherosclerosis risk in communities (ARIC) study. Hypertension 42(6):1106–1111

Song Y, Ni H, Zhou X, Zhao W, Wang T (2015) Extracting Features for Cardiovascular Disease Classification Based on Ballistocardiography. In: 2015 IEEE 12th intl conf on ubiquitous intelligence and computing and 2015 IEEE 12th intl conf on autonomic and trusted computing and 2015 IEEE 15th intl conf on scalable computing and communications and its associated workshops (UIC-ATC-ScalCom), pp 1230–1235. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.223

Van Oosterom A (2009) Measuring the T wave of the electrocardiogram; the how and why. Measur Sci Rev 9(3):53

Wang T, Wang Z, Zhang D, Gu T, Ni H, Jia J, Zhou X, Lv J (2016) Recognizing Parkinsonian Gait pattern by exploiting fine-grained movement function features[J]. ACM Trans Intell Syst Technol 8(1):6

Wen W (2012) The status quo and countermeasures of hypertension prevention and treatment in China[J]. J Med Res 41(5):3–5

Wiens AD, Johnson A, Inan OT (2017) Wearable sensing of cardiac timing intervals from cardiogenic limb vibration signals. IEEE Sens J 17:1463–1470. https://doi.org/10.1109/JSEN.2016.2643780

Xiaonan G, Jian L, Cong S, Hongbo L, Yingying C, Mooi CC (2018) Device-free personalized fitness assistant using WiFi. Proc ACM Interact Mob Wearable Ubiquitous Technol 2(4):1–23. https://doi.org/10.1145/3287043 (**Article 165**)

Zheng Y, Yan BP, Zhang Y, Yu CM, and Poon CCY (2013) Wearable cuff-less PTT-based system for overnight blood pressure monitoring. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp 6103–6106. https://doi.org/10.1109/EMBC.2013.6610945

Zhou Z (2016) Machine learning. Tsinghua University Press, Beijing (**ISBN 978-7-302-42328-7**)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.