Computer-Aided Atrial Fibrillation Diagnosis System with the Naive Bayesian Network

Computer-Aided Atrial Fibrillation Diagnosis System with The Naive Bayesian Network: Based on The Analysis of 2016 Actual Cases of Electrocardiography Signals

Conference Paper · January 2022

DOI: 10.1109/ICCECE54139.2022.9712791

3 authors, including:

Yuhao Sun
The University of Edinburgh
8 PUBLICATIONS **19** CITATIONS
SEE PROFILE

Junsheng Sun
□□□□□□□□
5 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Computer-Aided Atrial Fibrillation Diagnosis System with The Naive Bayesian Network: Based on The Analysis of 2016 Actual Cases of Electrocardiography Signals

Yuhao Sun
The Roslin Institute
The University of Edinburgh
Edinburgh, United Kingdom
yuhao.sun@ed.ac.uk

Yamei Zhao
Department of Rehabilitation
Longgang Central Hospital of Shenzhen
Shenzhen, China
yameizhao@126.com

Junsheng Sun*
Department of General Practice
Longgang Central Hospital of Shenzhen
Shenzhen, China
* Corresponding author:
junshengsun@126.com

2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE) | 978-1-6654-0886-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCECE54139.2022.9712791

***Abstract*—Atrial Fibrillation (AF) is a form of arrhythmia that occurs often. Over one million people in the UK have been diagnosed with AF, which may result in the development of other severe disorders, eventually posing a risk to life. We developed a computer-aided approach for AF diagnosis based on the Naive Bayesian Network based on 2016 real electrocardiogram (ECG) signal cases in order to determine whether the candidate was healthy or has AF. Accuracy was up to 97%. Due to the simplicity of this technology, it potentially offers a low-cost solution for areas that cannot afford a costly AF diagnostic system. This paper details the whole process of the research, from its inception to its conclusion, including the first thoughts and the related work, the methodology, and the discussion. It also primarily demonstrates the program's technique, including the program's unique probabilistic reasoning process and the 10-fold cross-validation testing results.**

Keywords-Atrial Fibrillation (AF); Naive Bayesian Network; Electrocardiography (ECG); Computer-Aided Diagnosis

I. INTRODUCTION

Atrial Fibrillation (AF) is a form of arrhythmia that occurs often. Over a million people in the UK and 90 million people worldwide have been diagnosed with AF, and those with AF may face a fivefold increased risk of stroke [1]. By 2050, the number of AF patients is expected to rise to 6–16 million in the United States and 72 million in Asia [2]. Additionally, AF is difficult to detect and frequently goes misdiagnosed. Most significantly, AF has strong epidemiological connections with a variety of other disorders, including diabetes mellitus, hypertension, cardiac valve disease, obesity, metabolic syndrome, and sleep apnea [3]. As a result, hospitals and healthcare facilities are highly recommended to have an effective AF diagnostic system. We developed a unique system based on 2016 examples of 12-lead electrocardiogram (ECG) signals using one of the probabilistic reasoning methods—Naive Bayesian Network. The data was retrieved from [4]. A five-step approach was developed, consisting of the following steps: input raw ECG data, data preprocessing, feature extraction, naive Bayesian network construction, and output result. With the appropriate testing, the method we built achieved 97% accuracy. The method potentially provides a low-cost alternative to pricey AF diagnostic equipment for areas that cannot afford it. Additionally, we have discussed reasons for choosing the Naive Bayesian Network as the primary technique. This article details the whole research process, from the initial thoughts through the methodology, discussion, and conclusion.

II. INITIAL CONSIDERATIONS

Prior to beginning the design and implementation of the research, it was good to consider numerous basic features, including but not limited to AF and ECG, in order to have a better understanding of the subject. The initial concepts incorporate current clinical viewpoints, a current literature review of computer-aided AF diagnosis system, understanding of AF and 12-lead ECG, and various ways of probabilistic reasoning in this context.

A. Clinical Perspectives

Because each sample provides 12-lead ECG data, it is critical to identify the core data among the 12-lead to reduce programme load and increase performance effectively. Dr Junsheng Sun, head of the Department of General Practice in a general tertiary hospital in Shenzhen China, said, “Typically, we evaluate the numbers generated by all 12-leads with an emphasis on the critical leads, such as the II-lead and V1-lead. Additionally, several elements are visible within a single figure, most notably the P wave, QRS wave, RR interval, and F wave, among others. However, diagnosing AF should consider all probable circumstances, otherwise it will result in a mistake. For instance, one of the characteristics of Second-degree Atrioventricular (AV) Block Type I is a different RR interval, which is also the primary characteristic of AF.”

B. AF and ECG

ECG is a frequently used technique for diagnosing AF. According to ECG indications, a probable AF patient may have

978-1-6654-0886-8/22/$31.00 ©2022 IEEE

a faster cardiac rate than normal persons, or an irregular rhythm. Additionally, some situations, such as the absence of P waves, the absence of an isoelectric baseline, changing ventricular rates, and the presence of fibrillatory waves, might possibly target AF patients.

C. Related Work

Currently, ECG equipment offers a variety of sophisticated capabilities that may alert clinicians to suspected AF patients immediately while also providing an initial diagnostic report. To achieve the powerful functionalities, many deep learning approaches have become popular [5-9]. Additionally, automated identification of AF using probabilistic reasoning approaches has made significant progress. Martis developed a system for ECG-based pattern analysis of normal sinus rhythm and atrial fibrillation (AF) beats in 2013 by combining Naive Bayes and Gaussian Mixture Model (GMM) [10]. Li also assesses the prevalence of AF using a three-class mixture model with the Expectation-Maximization (EM) technique for repeated diagnosis [11]. Several further references cited Bayes Classifier [12] and EM [13] while developing a method for identifying AF patients. According to a review, five major stages have typically been involved in the creation of a Computer-Aided Diagnosis (CAD) system for AF diagnosis: "Input ECG signals", "Pre-processing", "Feature Extraction", "Feature selection/ranking", and "Classification" [1].

D. Probabilistic Reasoning in AF

Numerous probabilistic reasoning approaches might be explored for developing a system for diagnosing AF patients in this context, including Naive Bayesian Classifier, Bayesian Classifier, Expectation Maximization, and Gaussian Mixture Model.

Among the probabilistic reasoning methods discussed above, the Naive Bayesian Network is an ideal choice in this scenario, with the following benefits, particularly considering the current situation:

- It is straightforward and simple to construct theoretically and has the potential to function efficiently.
- It requires less training data. 2016 cases in our research may be considered a small number in comparison to the data required for artificial neural networks (ANN) or convolutional neural networks (CNN).
- It is quick and may be used to generate real-time predictions.

III. Methodology

The research methodology was composed of five major steps: input (raw ECG data), preparation, feature extraction, the Naive Bayesian Network, and output (AF or normal). Figure 1 graphically illustrates the methodology's design. The program was developed on MATLAB. As aforementioned, all ECG data in our research was retrieved from an open access database [4].

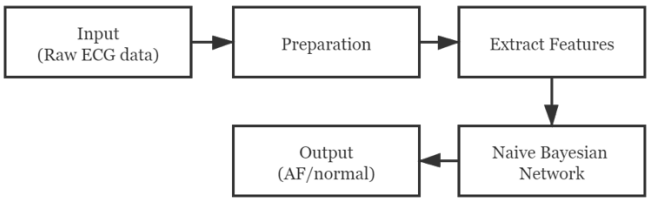

Figure 1. Methodology.

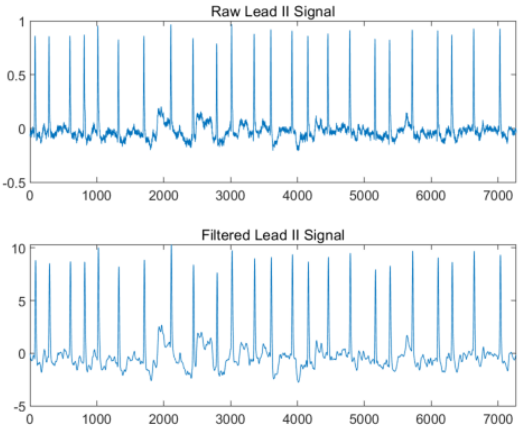

Figure 2. Raw and Filtered Signals of "A0126".

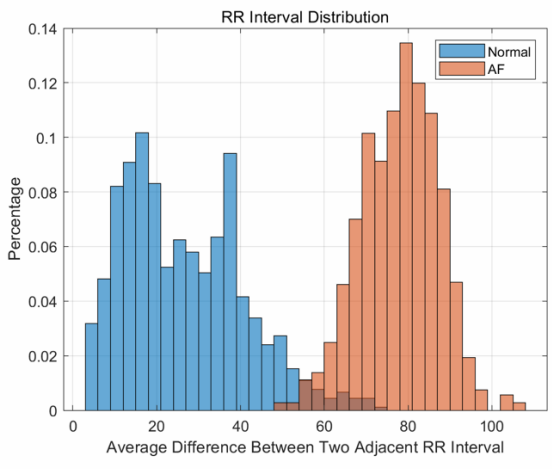

Figure 3. RR Interval Difference Distribution.

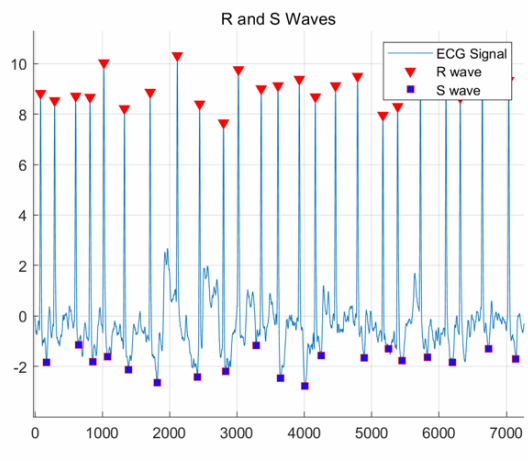

Figure 4. R and S Waves of The Case "A0126".

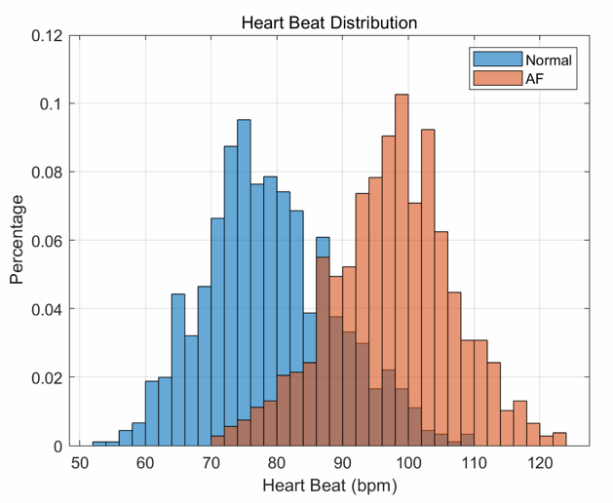

Figure 5. Heartbeat Distribution.

A. Preparation

The first step of the methodology was the necessary preparation work, including cleaning the workstation (i.e. MATLAB), loading the data file into the workstation, and filtering signals. A function conv() , with a kernel for the low pass filter, [1 1 1 1 1 1 1 1 1 1 1 1 1 1 1] has been used as filtering the signals. A comparison between raw and filtered signals please refer to Figure 2.

Please notice that the II-lead was merely loaded as it has been considered as one of the most critical leads in the clinical domain. Other leads are also valuable; but, under this research, one of the 12 leads was adequate to determine if the patient has AF or not. Other specialised leads, such as V1-lead, could also be used; however, rigorous testing revealed no significant difference in the outcomes.

B. Extract Features

The second step was to extract characteristics from ECG data that could be utilised as diagnostic factors for AF. Along with the early concepts, another program was designed to identify valuable qualities at the start. Three characteristics were finally adopted: the RR interval, the pulse, and the R peak. The next section addresses each of them in-depth.

1) RR Interval: The RR interval was the first characteristic that has been accepted. Within the same time, typical people's RR intervals (for example, the previous and next RR intervals) should be same or similar to the maximum extent. The variations in RR intervals between normal persons and AF sufferers can be shown graphically as two hills—or two Gaussian distributions—in Figure 3. Please note that the numbers on the X-axis do not represent the values of the RR intervals; rather, they represent the average value of the difference between all previous and subsequent RR intervals within the specified ECG period. Prior to computation, all values were converted to their absolute values. To determine the average difference between RR intervals, the values of the RR intervals must first be calculated. The values of the R peak were determined using the findpeaks() function in MATLAB with the "MinPeakHeight" and "MinPeakDistance" parameters set to 0.1 and 200, respectively, as shown in Figure 4. Please take note that the initial peak has been deleted due to its outlier status. The RR intervals were determined using the R peak values by subtracting the two corresponding x values of two neighbouring R peaks. Following that, the difference between two consecutive RR periods was computed. For instance, if a sample comprises ten peaks, it should have nine RR intervals and eight RR interval differences. The final outcome of this step was the average value of the sum of all RR interval differences for each sample. As seen in Figure 3, a lower number on the x-axis indicates a reduced likelihood of AF, and vice versa. The result derived from this characteristic was the average value of the RR interval difference for each sample.

2) Heartbeat: The second characteristic that has been incorporated was the heartbeat. The heartbeats of normal individuals and AF patients can also be shown graphically as two hills—or two Gaussian distributions—in Figure 5. The values on the x-axis represent the average beats per minute (bpm) of the provided sample during the specified time period. At this stage, it was easier to find a heartbeat value, because RR intervals were received by the earlier step. As all ECG data has been sampled at 500 Hz, bpm could be therefore calculated by:

```
sec_per_beat = RR_avg * 1 / (500);
bpm = 60 / sec_per_beat;
```

The result retrieved from this feature was the average value of the heartbeat for each sample.

3) R Peak: The R peak was the third and final characteristic that has been accepted. The values of R peak in normal persons and AF patients may alternatively be graphically represented as two hills—or two Gaussian distributions—in Figure 6. However, certain areas of the two distributions overlapped and were not as different as the two qualities mentioned above. Additionally, please notice that the numbers on the x-axis do not represent the value of the R peak; rather, they represent the

average value of the difference between all prior and subsequent R peaks within the specified ECG period (similar to calculating the first feature in Section III-B1). Prior to computation, all values were converted to their absolute values. Section III-B1 determined the R peak values. The difference between the previous and subsequent R peaks were also determined in this manner. Thus, the result derived from this characteristic was the average value of the R peak difference for each sample.

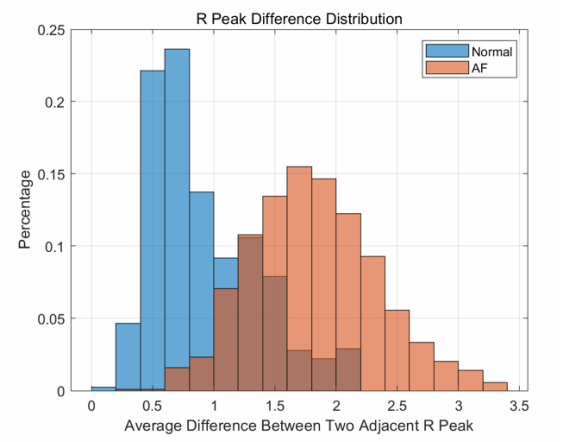

Figure 6. R Peak Difference Distribution.

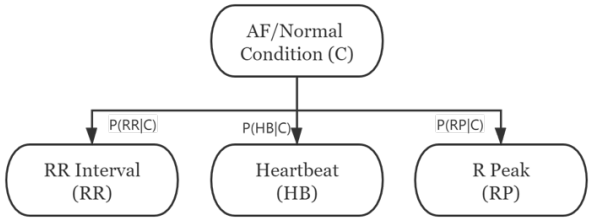

Figure 7. Link Matrix of Naive Bayesian Network.

C. Naive Bayesian Network

It was the core to build a suitable Naive Bayesian Network, which was the probabilistic reasoning technique employed in our research. Section II-D outlined the advantages of adopting Naive Bayesian Networks in the current environment. The connection matrix of the planned Naive Bayesian Network is shown in Figure 7. A sequence of processes must be followed in order to get at the categorisation. Additionally, please notice that the numbers below correspond to the block index in the source codes.

1) Preparation: With the features extracted from the earlier step, the preparation work was implemented in this step. Firstly, to easily find the data of features, we constructed a matrix which contains all features, including AF/normal conditions. Variable named as "all_data", in which the 1st row was the index of samples (1:918 belongs to normal samples and 919:2016 belongs to AF patients); the 2nd row was AF/normal conditions (1:918 was 0 meaning no AF and 919:2016 was 1 meaning AF patients); the 3rd row was RR interval difference (variable named as "rr"); the 4th row was heartbeat value (variable named as "hb"); and the 5th row was R peak difference (variable named as "rp"). Additional grouping work should be performed for the 10-fold cross-validation. This entails separating all samples (in this case, the first 2000 samples) into ten divisions, each of which has 200 samples and was treated as separate testing set for each time. Meanwhile, 1800 samples, i.e. the remaining 9 divisions, were used as the training set. Ten testing results were created iteratively and discussed in depth in Section IV-A.

2) Calculation of Gaussian Distribution: As previously proven, the derived characteristics' data were not binary, i.e. yes or no; rather, they were all continuous variables. Thus, for each characteristic, the data could be seen as two Gaussian distributions (AF's and normal's) and the appropriate mean and standard deviation values could be determined for each distribution. The chance of each instance could be determined using the normal probability density function and the aforementioned mean and standard deviation. In MATLAB, normpdf() is a handy function for implementing the preceding concept. Additionally, the number of AF patients and normal samples has been estimated across all input data (especially in testing cases, both were not 918 and 1098 respectively anymore). "af_counter" and "normal_counter" indicated the number of AF patients and normal samples, respectively.

3) Calculation of Probability: The conditional probability was calculated within this step. Firstly, each conditional probability was calculated based on the provided training data, which included P(RR|yes), P(RR|no), P(HB|yes), P(HB|no), P(RP|yes) and P(RP|no). Specifically, P(RR|yes) and P(RR|no) stand for the conditional probability of the specific RR interval difference when the sample is AF and normal respectively, so on and so forth. Secondly, the probability of AF patients and normal samples among all input data was also be calculated. This could be simply done by using divisions of "af_counter/2016" and "normal_counter/2016", in which there were 2016 input samples (for 10-fold cross validation, this should be 1800).

4) Classification: In the end, the classifying work was completed, according to (1) and (2).

$$P(yes) = \alpha \times P(af) \times P(RR|yes) \times P(HB|yes) \times P(RP|yes) \quad (1)$$

$$P(no) = \alpha \times P(norm) \times P(RR|no) \times P(HB|no) \times P(RP|no) \quad (2)$$

The value of alpha was calculated by 1/(P(yes)+P(no)) with temporarily removing alpha when calculating P(yes) and P(no) at the first time. With the new value of alpha, P(yes) and P(no) should be re-calculated by the full equations shown above. The sum of P(yes) and P(no) should be exactly 1. Two classification types follow:

- When P(yes) is larger than P(no), then the system results in a 1, meaning "AF" diagnosis.

- When P(no) is larger than P(yes), then the system results in a 0, meaning "normal" (healthy) diagnosis.

This is the end of Naive Bayesian Network. By extending the flowchart Figure 1 above, a thorough flowchart of the whole methodology can refer to Figure 8.

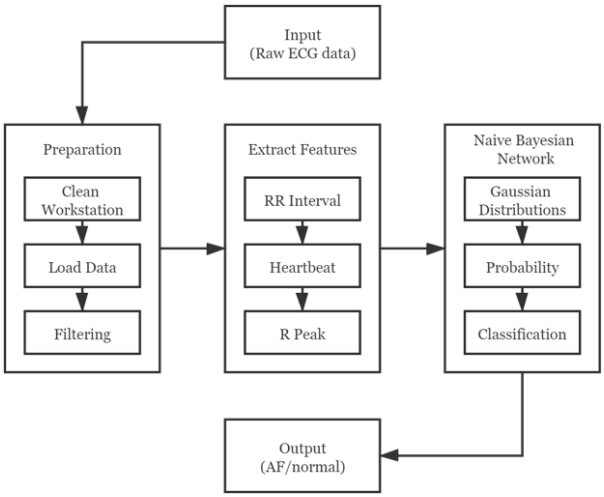

Figure 8. Full Methodology in Summary.

IV. DISCUSSION

A. Testing

Testing was conducted after the construction of the above system. As there were 2016 samples of ECG data totally, to strictly execute 10-fold cross validation, only the first 2000 data were used when testing. Each testing set contained 200 samples of data while the corresponding training set contained left 1800 samples of data. 10 testing results shown in Table I stand for the testing data sequence as (1:200), (201:400), (401:600), (601:800), (801:1000), (1001:1200), (1201:1400), (1401:1600), (1601:1800), and (1801:2000), respectively. For each testing case, a "resultlist" was created and used for calculating the accuracy compared with the provided known results.

Table I demonstrates the testing results to 10-fold cross validation. The mean and standard deviation of accuracy have been also shown in the table. Testing results show a positive performance to the Naive Bayesian Classifier constructed earlier. According to testing results, the program can deduce if a subject is normal or AF from 12-lead ECG signals with an average 97% accuracy which can be deemed as an excellent performance. Additionally, a relatively low standard deviation proves that the high accuracy calculated earlier has universality, at least for 2000 samples used.

TABLE I. 10-FOLD CROSS-VALIDATION RESULTS

Item	10-Fold Cross Validation				
Accuracy (%)	**1**	**2**	**3**	**4**	**5**
	98	90.5	98.5	90	98.5
	6	**7**	**8**	**9**	**10**
	99.5	99.5	99.5	100	96
Mean Accuracy (%)	97				
Standard Deviation of Accuracy	0.0373				

B. Findings

By viewing the Figure 3, 5 and 6, it is not hard to ask such a question: why not the doctors just view one of the features and give an initial diagnosis—especially in Figure 3, the RR interval difference of the normal people and AF patients have almost already been grouped by two Gaussian distributions. However, the classifier designed in this research has decreased the error to 3%; even though this is still high compared to some up-to-date ECG machines clinical using (which can have an error rate less than 1%), while it is better than merely using eyes. The classifier in the research is not as complicated as the latest machines but with a comparable result.

Thus, to create high-performance ECG machines with necessary probabilistic reasoning methods possibly can decrease the cost of medical equipment/software—not as expensive as those which adopt neural networks or more complicated artificial intelligence techniques. Some undeveloped areas and countries where lack of medical machines possibly could be beneficial from this idea.

C. Limitations

There are some limitations in this research. Potentially, several limitations exist mainly around the probabilistic reasoning method adopted and potential circumstances:

1) The quantity of characteristics is insufficient: Three characteristics (RR interval, heartbeat, and R peak) were used in the research. Naive Bayesian is a conditional probability model, which indicates that the more potential characteristics, the more exact the outcome.

2) Possibility of suboptimal performance when confronted with vast input data: At the outset, 2016 ECG data samples were presented. This is not a large dataset. However, if the input data size increases to ten thousand, one hundred thousand, or even one million, the Naive Bayesian Network becomes inefficient. This is a shortcoming of the Naive Bayesian approach in general.

3) "Are all features really independent?": The naive Bayesian approach implies that all predictors (or variables) are independent, which is extremely unlikely in practise. A critical question is: to what extent are they self-contained? For instance, the heartbeat may have an effect on RR intervals. Additionally, RR intervals may be determined using R peaks—what is the link between R peak and RR interval? Although this may need

further and deeper medical expertise, the answers to these questions will have an effect on the efficiency of the Naive Bayesian Network, either favourably or adversely.

4) Inadequate medical knowledge: This was a significant hurdle at the outset of the project. Thus, it demonstrates the importance of the aforementioned initial considerations. We are willing to believe that more collaboration with clinical staff will make the research more in line with the needs of them.

V. Conclusion

In conclusion, based on the initial thoughts summarised, a high-performance program has been successfully constructed to determine whether a potential candidate is normal or AF by properly utilising Naive Bayesian Networks with three features of RR interval, heartbeat, and R peak extracted from provided ECG data. A five-stage technique—input, preparation, extract features, Naive Bayesian Network, and output—has been created and introduced previously, along with the specifics for each phase. Additionally, the findings of 10-fold cross-validation effectively has demonstrated the developed system's excellent performance with a 97% accuracy. Finally, the research's findings and limitations have been reviewed, and all of which could be carefully considered in future study. Although the suggested approach is not sophisticated, it has the potential to serve as a viable alternative in underdeveloped areas that cannot afford expensive AF diagnostic equipment or software.

Acknowledgment

The research is supported by Longgang Central Hospital of Shenzhen, China. We also appreciate the clinical perspectives provided by Dr Junsheng Sun throughout this research. The source codes of this research can be accessed upon request.

References

[1] Y. Hagiwara *et al.*, "Computer-aided diagnosis of atrial fibrillation based on ECG signals: A review," *Information Sciences,* vol. 467, pp. 99-114, 2018.

[2] J. Kornej, C. S. Börschel, E. J. Benjamin, and R. B. Schnabel, "Epidemiology of atrial fibrillation in the 21st century: novel methods and new insights," *Circulation research,* vol. 127, no. 1, pp. 4-20, 2020.

[3] G. Lippi, F. Sanchis-Gomar, and G. Cervellin, "Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge," *International Journal of Stroke,* vol. 16, no. 2, pp. 217-221, 2021.

[4] F. Liu *et al.*, "An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection," *Journal of Medical Imaging and Health Informatics,* vol. 8, no. 7, pp. 1368-1373, 2018.

[5] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Systems with Applications,* vol. 115, pp. 465-473, 2019.

[6] M.-Z. Poh *et al.*, "Diagnostic assessment of a deep learning system for detecting atrial fibrillation in pulse waveforms," *Heart,* vol. 104, no. 23, pp. 1921-1928, 2018.

[7] S. Kwon *et al.*, "Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: algorithms development study," *JMIR mHealth and uHealth,* vol. 7, no. 6, p. e12770, 2019.

[8] J. Wang, "A deep learning approach for atrial fibrillation signals classification based on convolutional and modified Elman neural network," *Future Generation Computer Systems,* vol. 102, pp. 670-679, 2020.

[9] F. Murat *et al.*, "Review of Deep Learning-Based Atrial Fibrillation Detection Studies," *International journal of environmental research and public health,* vol. 18, no. 21, p. 11302, 2021.

[10] R. J. Martis, U. R. Acharya, H. Prasad, C. K. Chua, and C. M. Lim, "Automated detection of atrial fibrillation using Bayesian paradigm," *Knowledge-Based Systems,* vol. 54, pp. 269-275, 2013.

[11] L. Li, H. Mao, H. Ishwaran, J. Rajeswaran, J. Ehrlinger, and E. H. Blackstone, "Estimating the prevalence of atrial fibrillation from a three‐class mixture model for repeated diagnoses," *Biometrical Journal,* vol. 59, no. 2, pp. 331-343, 2017.

[12] Y. Ebana and T. Furukawa, "Networking analysis on superior vena cava arrhythmogenicity in atrial fibrillation," *IJC Heart & Vasculature,* vol. 22, pp. 150-153, 2019.

[13] S. Ladavich and B. Ghoraani, "Developing an atrial activity-based algorithm for detection of atrial fibrillation," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2014: IEEE, pp. 54-57.