IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Combining convolutional neural network and distance distribution matrix for identification of congestive heart failure

Yaowei Li[1#], Yao Zhang[2#], Lina Zhao[3#], Yang Zhang[4#], Chengyu Liu[1*], Li Zhang[5], Liuxin Zhang[4], Zhensheng Li[4], Binhua Wang[6], EYK Ng[7], Jianqing Li[1], Zhiqiang He[4*]

[1]The State Key Laboratory of Bioelectronics, Jiangsu Key Lab of Remote Measurement and Control, School of Instrument Science and Engineering, Southeast University, Nanjing, China
[2]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
University of Chinese Academy of Sciences, Beijing, China
[3] School of Control Science and Engineering, Shandong University, Jinan, China
[4] Lenovo Research, Beijing, China
[5] Computational Intelligence Group, Northumbria University, Newcastle upon Tyne, UK
[6] Medical Big Data Center, Chinese PLA General Hospital, Beijing, China
[7] School of Mechanical and Aerospace Engineering, College of Engineering, Nanyang Technological University, 639798, Singapore

Co-first authors

Corresponding author: Chengyu Liu (e-mail: chengyu@seu.edu.cn), Zhiqiang He (email: hezq@lenovo.com)

This work was partly supported by the National Natural Science Foundation of China (61571113 and 61671275), the Key Research and Development Programs of Jiangsu Province (BE2017735) and the Fundamental Research Funds for the Central Universities (2242018k1G010). The authors thank the support from the Southeast-Lenovo Wearable Heart-Sleep-Emotion Intelligent Monitoring Lab.

ABSTRACT Congestive heart failure (CHF) is a serious pathophysiological condition with high morbidity and mortality, which is hard to predict and diagnose in early age. Artificial intelligence and deep learning combining with cardiac rhythms and physiological time series provide a potential to help with solving it. In this study, we proposed a novel method that combines convolutional neural network (CNN) and distance distribution matrix (DDM) in entropy calculation to classify CHF patients from normal subjects, and demonstrated the effectiveness of this combination. Specifically, three entropy methods were used to generate the distribution matrixes from a 300-point RR interval (i.e., the time interval between the successive cardiac cycles) time series, which are Sample entropy (SampEn), fuzzy local measure entropy (FuzzyLMEn) and fuzzy global measure entropy (FuzzyGMEn). Then, three high representative CNN models, i.e. AlexNet, DenseNet and SE_Inception_v4 were chosen to learn the pattern of the data distributions hidden in the generated distribution matrixes. All data used in our experiments were gathered from the MIT-BIH RR Interval Databases (http://www.physionet.org). A total of 29 CHF patients and 54 normal sinus rhythm (NSR) subjects were included in this study. The results showed that the combination of FuzzyGMEn-generated DDM and Inception_v4 model yielded the highest accuracy of 81.85% out of all proposed combinations.

INDEX TERMS Congestive heart failure (CHF), convolutional neural network (CNN), distance distribution matrix (DDM), heart rate variability (HRV), entropy

I. INTRODUCTION

Congestive heart failure (CHF) is a serious pathophysiological condition, which has become a common cause of hospitalization with significant morbidity and mortality [1-4]. However, heart failure remains insufficiently diagnosed worldwide, especially in early age [5-8]. Precise diagnosis is thus vital for heart failure treatment. Previous studies showed that heart rate variability (HRV), which is associated with the mortality of CHF, is an effective feature for discriminating CHF patients from normal subjects [9-11]. Over the past years, various machine learning methods were proposed to diagnose patients suffering from CHF based on HRV. For example, Isler et al. proposed a model based on k-nearest neighbor classifier (KNN) and wavelet entropy [12]. Jovic et al. utilized random forest and combinations of linear and nonlinear features of HRV [13]. Pecchia et al. designed a classifier based on regression tree with selected RMSSD, total power, HF, and LF/HF as useful classification features [11]. There are also researchers who employed SVM and combinations of several HRV features and achieved relatively high accuracies [14-16].

Most existing works employ classifiers with comparatively simple structures and trained on small data sets. The input of their classifiers is empirically a set of selected features. However, the performance of the classifiers is largely based on feature selection processes [12, 14]. Thus in most cases, a large amount of time and effort is paid to manually find better feature subsets and even adopted the so-called exhaustive search methods to find the best subsets of features [17]. Additionally, the choice of the best feature combination may change with different datasets. With the explosion of data and the development of smart wearable devices, deep learning is a desirable way to overcome the shortage of artificial feature extraction and selection. Deep neural networks are designed to automatically learn the underlying hidden feature combinations without any manual process. As one type of the most successful deep neural network, convolutional neural network (CNN) has gained significant development and achieves state-of-the-art results on various tasks [11]. CNNs are able to accept raw and complete images as inputs, so as to avoid the risk of losing valuable information. Thus, we decide to employ different CNNs to automatically learn effective features from HRV data and produce accurate classification results without complicating manual feature extraction.

Entropy is a non-linear HRV analysis method, which provided a better understanding for the underlying mechanisms of the cardiovascular system [18-20]. In previous study, entropy calculation was able to distinguish CHF and normal sinus rhythm (NSR) subjects with appropriate parameters. A statistical significance for the two groups was obtained [21, 22]. Jovic et al. tried to use combinations of entropy calculation results as the input of classifiers and acquired a moderate result of approximate 73% accuracy [13]. It could be attributed to the simple and rough entropy calculation, i.e. there will be only a number value result, leading to a potential risk to lose useful information for subsequent normal/abnormal classification.

The construction of distance distribution matrix (DDM) is an essential step for entropy calculation. The difference between normal and abnormal cardiac conditions can be depicted and observed by DDM. This is thus a desirable input for CNN as it reveals the features of HRV signals in the manner of entropy analysis but contains richer information than a simple single entropy value calculation. The RR interval is the time interval between the successive cardiac cycles and regarded as an important feature of an ECG signal. It is usually quantified by the time difference between the occurrence of the maximum wave, i.e. the R wave of a cardiogram. Thus RR interval time series in the long-term RR Interval Databases from http://www.physionet.org [23] are used in this study to generate the DDMs.

In this study, our main aim is to use the DDM as an image feature to achieve classification between the NSR and CHF subjects by employing these improved representative CNN methods. Several stages were included in this study. The first stage is to convert RR interval time series into DDMs using three kinds of entropy methods: i.e. Sample entropy (SampEn), fuzzy local measure entropy (FuzzyLMEn) and fuzzy global measure entropy (FuzzyGMEn). The second stage is to train classifiers based on three different types of CNN models. Experimental study is presented in the last stage, which evaluates our models on two schemes. Our contributions are summarized as follows:

1) We improve three different types of classifiers without manual feature extraction based on latest state-of-art CNN models.
2) We generate three kinds of DDMs from RR interval time series as the input of these classifiers and compare their classification results based on the three CNN classifiers. All three kinds of DDMs show discriminability for the RR interval time series between NSR and CHF groups, and the performance of each model has no significant difference. This verifies the effectiveness of combination of DDM and the CNN model.
3) We choose the subject-based and segment-based schemes as the evaluation schemes and compared their performances. In this study, the segment-based scheme performs similarly to the subject-based scheme.

II. CNN MODELS

AlexNet [24], DenseNet [25] and Inception_v4 [26] were used in this study. AlexNet is one of the largest CNNs trained on the subsets of ImageNet used in the ILSVRC-2010 and ILSVRC-2012 competitions. DenseNet alleviates the disappearance of gradients and enhances feature propagation by encouraging feature reuse, and this greatly reduces the amounts of parameters. Inception_v4 was one of several follow-up versions to GoogLeNet [27], and is the winner of ILSVRC 2014, but became deeper and wider by introducing residual connections and has a more simplified architecture and more inception modules than the previous versions [26]. All these models are representative CNN models. The details of the three employed CNNs as described as follows:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2855420, IEEE Access

A. ALEXNET

Original AlexNet contains five convolutional and three fully-connected ones. In our study, we converted those fully-connected layers into convolutional layers. This made it possible to efficiently run the CNN on 297×297 input images. The architecture was summarized in Fig. 1. Firstly, we use a convolution with 64 output channels and kernel size 11×11 to input distribution matrix followed by a 3×3 max pooling layer. After several convolution and max pooling operations, dropout layers were also used to enhance the robustness of the model. At the end of the network, the global average pooling layer is performed. Besides, rectified linear units (ReLUs) were used to reduce training time and local normalization scheme was used to aid generalization.

B. DENSENET

DenseNet consists of alternating transition layers and dense blocks. Fig. 2 illustrates the architecture of the DenseNet. Firstly, we use a convolution with 48 output channels followed by a transition layer. Each transition layer is to change the size of feature maps by convolution and pooling between dense blocks, which consists of a batch normalization layer, a ReLU layer and a 1×1 convolutional layer with 24 output channels followed by a 2×2 average pooling layer. In a dense block, each layer obtained additional inputs from all its preceding layers and passes on its own feature maps to all its subsequent layers. The network is divided into multiple densely connected dense blocks.At the end of the DenseNet, a global average pooling is used and then a softmax classifier is performed.

C. INCEPTION-V4

The main contribution of Inception_v4 was the Inception Module that dramatically reduced the number of parameters in the network. Additionally, it used average pooling instead of fully connected layers at the top of the ConvNet, eliminating a large number of parameters without remarkably decrease of performance. In our study, we add “Squeeze-and-Excitation” (SE) block in each inception block to model channel-wise relationships in a computationally efficient manner. It enhance the representational power of modules throughout the network. Consequently, we term our model as SE_Inception_v4. The overview of SE_Inception_v4 is illustrated in the left side of Fig. 3. It is composed of “stem”, “inception” and “reduction” modules, as shown in Fig. 3 and Fig. 4 in detail.

FIGURE 1. **The architecture of AlexNet**

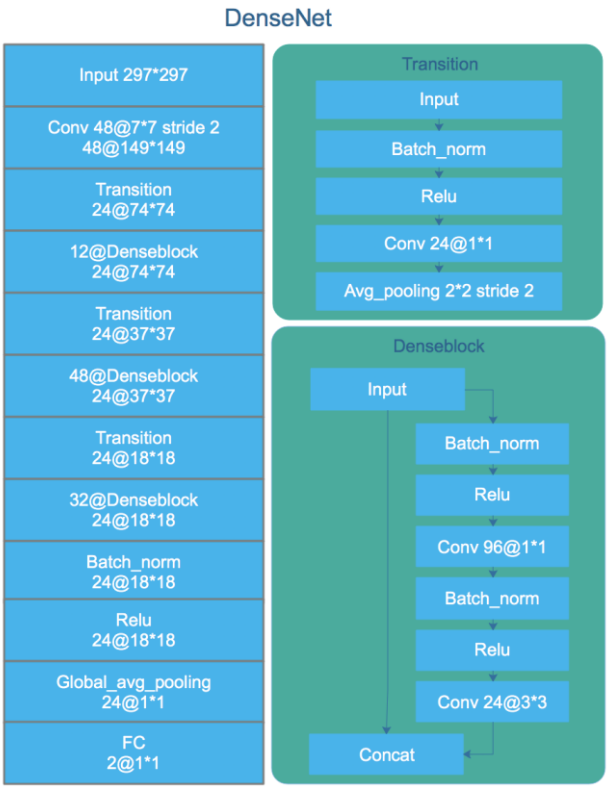

FIGURE 2. **The architecture of DenseNet**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2855420, IEEE Access

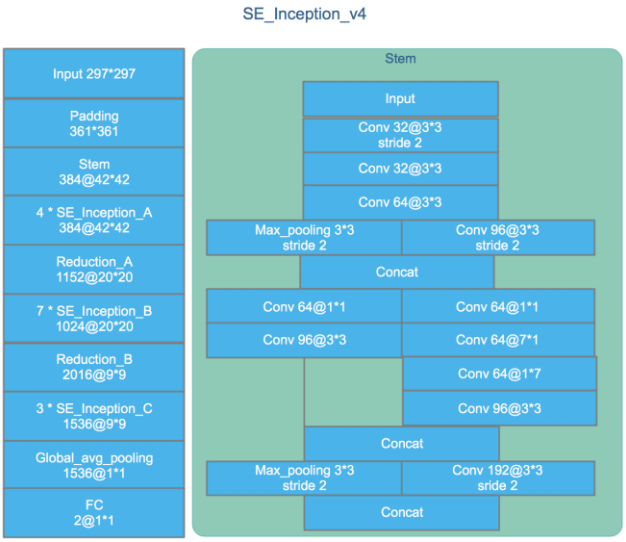

FIGURE 3. The whole architecture of SE_Inception_v4 and the "stem" module in SE_Inception_v4

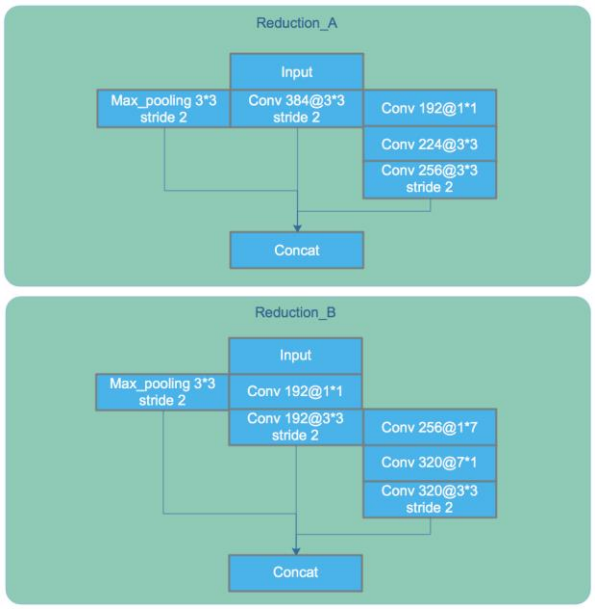

FIGURE 4. "Reduction" module in SE_Inception_v4

III. EXPERIMENT

A. DATA

All data used in our experiments were gathered from the long-term RR Interval Databases (http://www.physionet.org) [23], a free-access, on-line archive of physiological signals. The NSR RR Interval Database was used as the non-pathological and control group data. This database included 54 long-term RR

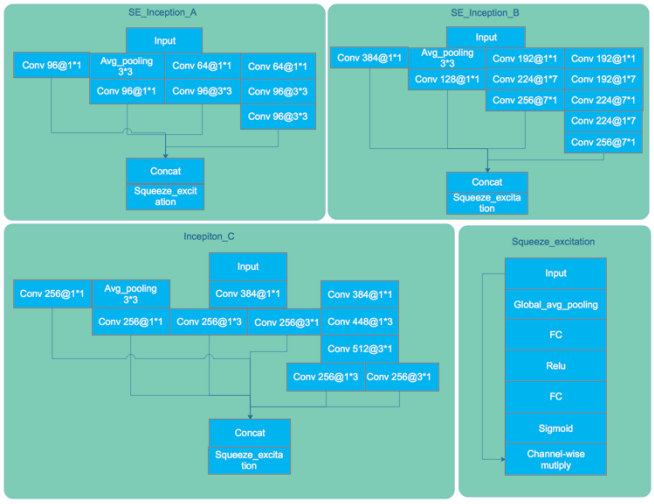

FIGURE 3. "Inception" and "Squeeze-and-Excitation" modules in SE_Inception_v4

interval recordings of subjects in normal sinus rhythm aged from 29 to 76. The CHF RR Interval Database was used as the pathological group data. This database included 29 long-term RR interval recordings of subjects aged from 34 to 79, with congestive heart failure (NYHA classes I, II, and III). The original ECG signals for both NSR and CHF RR interval databases were resampled at 128 Hz, and the beat annotations were obtained by automated analysis with manual review and correction.

B. PRE-PROCESS

RR interval is one of the important features of the ECG signal. It is the time interval between the successive cardiac cycles, which is usually quantified by the time difference between the occurrence of the maximum wave, R, of a cardiogram and is often called RR interval. In this section, two steps were used in the pre-process procedure for each RR interval recording:

Step 1: Each beat in the raw ECG signals was annotated as a normal or abnormal heartbeat. These abnormal heartbeats, usually caused by the ectopic beats such as supra-ventricular ectopic beats or ventricular ectopic beats (depending on the localization of the ectopic focus), were removed from the raw ECG signals, as the RR intervals formed from the abnormal heartbeats could confound the entropy analysis of HRV. We also remove RR intervals greater than 2 seconds to ignore the influence from the artifacts. Table 1 shows the total number of RR intervals for both NSR and CHF groups, as well as the numbers of RR intervals after the above procedure.

Step 2: Then we divide these ECG signals into several RR segments. The length of each RR segment is recorded as N, and we set N = 300, i.e. each RR segment contains 300 RR intervals.

TABLE I

STATISTICAL RESULTS OF THE NUMBERS OF RR INTERVAL RECORDINGS, RR INTERVALS AND RR SEGMENTS FROM THE 54 NSR AND 29 CHF RR INTERVAL DATABASES.

Variables	NSR group	CHF group
Name of RR interval recordings	NSR001-NSR054	CHF201-CHF229
No. of RR interval recordings	54	29
No. of RR intervals	5,790,504	3,312,195
No. of RR intervals after removing greater than 2s	5,780,148	3,306,394
No. of RR intervals after removing abnormal heartbeats	5,738,937	3,102,120
No. of RR segments when setting N=300	19,101	10,324

C. GENERATION OF DDM

SampEn [28], proposed by Richman and Moorman, can be used to analyze physiological time series [29]. SampEn quantifies the conditional probability that two sequences of m length similar consecutive data points will still be similar for m+1 (given a tolerance r). DDM generation is an intermediate step for SampEn calculation. DDM consists of similarity degrees which are determined by the distance and a decision rule. The distance is defined as follows:

For the HRV series $x(\mathrm{i}), 1 \le i \le N$, given the parameters m, form $N-m+1$ vectors

$$X_i^m = \{x(i), x(i+1), \cdots, x(i+m-1)\}$$
$$1 \le i \le N-m \quad (1)$$

The distance between any two vectors X_i^m and X_j^m based on the maximum absolute difference is defined as:

$$d_{i,j}^m = d\left[X_i^m, X_j^m\right] = \max_{k=0}^{m-1} \left|x(i+k) - x(j+k)\right| \quad (2)$$

where m denotes the embedding dimension.

The decision rule for vector similarity is based on the Heaviside function in SampEn. If the distance is within the threshold parameter r, the similarity degree between the two vectors is 1; if the distance is beyond the threshold parameter r, the similarity degree is 0. This rigid boundary may induce abrupt changes of entropy values when the tolerance threshold r changes slightly, and even fail to define the entropy if no vector-matching could be found [30-32]. To enhance the statistical stability, a fuzzy measure entropy (FuzzyMEn) method was proposed [31, 33], which used a fuzzy membership function to substitute the Heaviside function.

Unlike the 0 or 1 discrete determination for vector similarity degree in SampEn, fuzzy membership function permits the FuzzyMEn outputs continuous numerical values between 0 and 1 for the degree of vector similarity. Since FuzzyMEn not only measures the global vector similarity degree, but also refers to the local vector similarity degree. Thus, in this study we define FuzzyLMEn as the FuzzyMEn that is measured by local vector similarity degree. We also use FuzzyGMEn to denote the FuzzyMEn that is measured by global vector similarity degree. The detailed descriptions of SampEn, FuzzyLMEn and FuzzyGMEn were summarized in the Appendix.

Three types of DDMs are generated firstly at the setting of different embedding dimension m and m+1. Then we calculated the difference of these two DDMs. In the following classification process, the differences of DDMs were used as the input images of the CNN classifiers. Figures 6-8 show the DDMs generated by SampEn, FuzzyGMEn and FuzzyLMEn. We set embedding dimension m as 2 and 3 combined with threshold r = 0.1 and segment length N = 300, which has been proved statistical significance for SampEn, FuzzyGMEn and FuzzyLMEn [21]. Only $1 \le i \le 297$ and $1 \le j \le 297$ are shown for illustrating the details. In each sub-figure, the upper panel shows the results from a NSR subject, and the lower panel shows the results from a CHF subject. The results are from the embedding dimension m = 2, and m = 3 respectively. Their difference is showed from left to right respectively and are used as the input images of the CNN classifiers in the following classification process. Black colored areas indicate the similarity degree = 1 and vice versa.

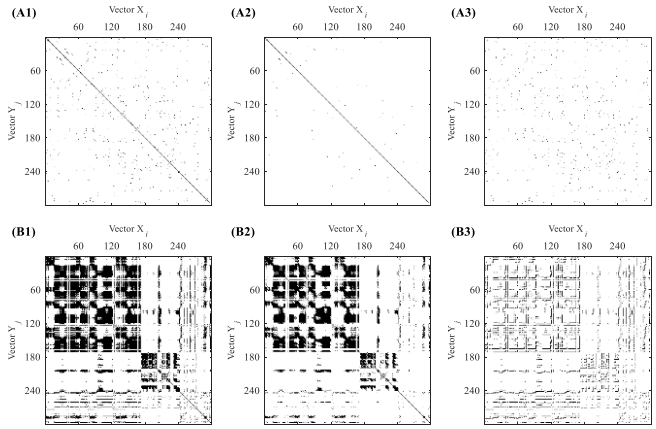

FIGURE 5. (A) DDM generated by SampEn for NSR subject under different parameter settings: (A1) m = 2, (A2) m = 3, (A3) the difference of (A1) and (A2); (B) DDM generated by SampEn for CHF patient under different parameter settings: (B1) m = 2, (B2) m = 3, (B3) the difference of (B1) and (B2).

Figure 6 presents the DDMs generated by SampEn. Figures 7-8 present the DDMs generated by FuzzyGMEn and FuzzyLMEn respectively. Unlike the 0 or 1 discrete determination for vector similarity degree in SampEn, FuzzyGMEn and FuzzyLMEn permit the outputs of continuous real values between 0 and 1 for the vector similarity degree, by converting the absolute distance of using a fuzzy exponential function (see Appendix). Dark-colored areas indicate the higher similarity degree and vice versa.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2855420, IEEE Access

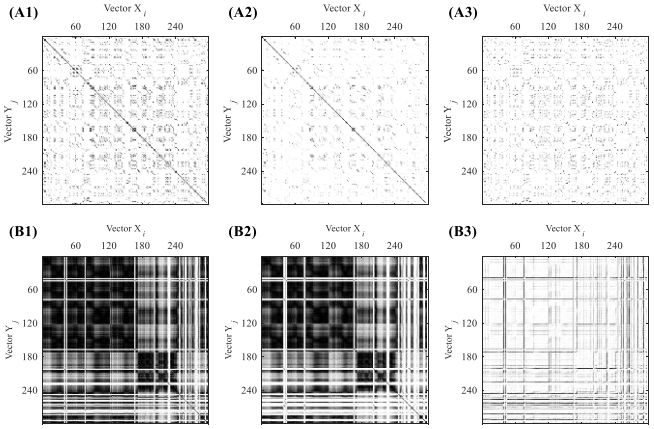

FIGURE 6. (A) DDM generated by FuzzyGMEn for NSR subject under different parameter settings: (A1) m = 2, (A2) m = 3, (A3) the difference of (A1) and (A2); (B) DDM generated by FuzzyGMEn for CHF patient under different parameter settings: (B1) m = 2, (B2) m = 3, (B3) the difference of (B1) and (B2).

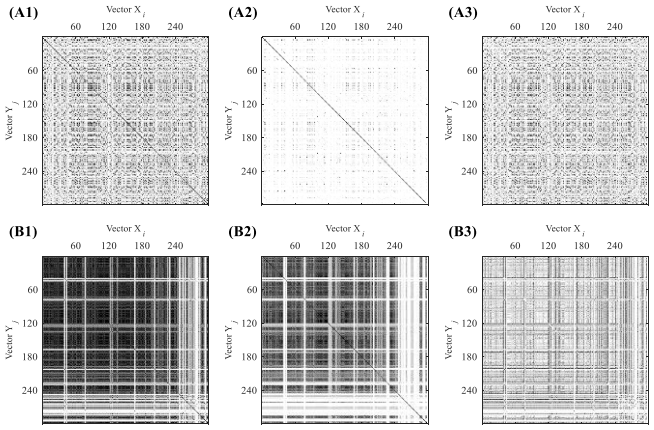

FIGURE 7. (A) DDM generated by FuzzyLMEn for NSR subject under different parameter settings: (A1) m = 2, (A2) m = 3, (A3) the difference of (A1) and (A2); (B) DDM generated by FuzzyLMEn for CHF patient under different parameter settings: (B1) m = 2, (B2) m = 3, (B3) the difference of (B1) and (B2).

D. MODEL CONFIGURATION

The details of AlexNet, DenseNet and SE_Inception_v4 are illustrated in Fig. 1, Fig. 2 and Fig. 3 respectively. All three models were implemented with Tensorflow library [34]. We trained the networks from scratch with a Gaussian random initializer ($\mu = 0$, $\sigma = 0.01$). The Adam optimizer with an initial learning rate of 0.0001 was used for parameters updating. The dropout was set to 0.5 to avoid overfitting.

E. EVALUATION SCHEME

In this study two schemes are considered for the selection of training and test sets. The first selecting scheme is based on subject (recording). We randomly select subjects into five folds. Four folds for training, and the remaining one is for testing. Table 2 shows the results of selecting.

TABLE II

FOLD RESULTS FOR ALL RECORDS IN THE TWO GROUPS

Fold#	CHF records	NSR records	Number	Number	Total
fold1	201,213,215,218	8,12,13,20,22,23,25,3	6	11	17
fold2	202,205,206,210	4,15,21,24,27,29,31,3	6	11	17
fold3	204,207,209,219	1,7,9,10,11,16,19,34,	6	11	17
fold4	203,216,217,221	5,6,17,26,28,32,35,42	6	11	17
fold5	208,211,212,214	2,3,14,18,30,33,36,40	5	10	15
total			29	54	83

Besides subject-based selecting scheme, we also consider segment-based scheme. To evaluate the robustness of the proposed models, 5-fold cross-validation strategy is employed. Firstly, the first 10% data of each subject are used to train and the other 90% of data are used to test without any overlap. Then the percent of train data increases by 10% and repeats until the first 90% data of each subject are used to train and the last 10% are used to test.

F. PERFORMANCE MEASURES

We evaluate our model performance by combining True/False Positives/Negatives to measure Precision, Recall and Accuracy (Acc.) [35]. They are often considered to be the most informative for characterizing the performance of a classifier and easy to calculate. Accuracy (Acc.) is the ratio of the total number of positives and negatives correctly made by the recognition system to the actual total number of positives and negatives confirmed by the recognition system. Precision measures the rate of true positives among all detections, while Recall measures the percentage of detected ground truth annotations. They are defined by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}},\ \text{Recall} = \frac{\text{TP}}{\text{TP+FN}},\ \text{Acc.} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (3)$$

where true positives (TP) denotes the number of CHF segments correctly classified as CHF group. False positives (FP) refer to the number of NSR segments incorrectly classified as CHF group. True negatives (TN) associate with the number of NSR segments correctly classified as NSR group. False negatives (FN) refer to the number of CHF segments incorrectly classified as NSR group.

IV. RESULTS

A. SUBJECT-BASED SCHEME

For the subject-based selecting scheme, Tables 3-5 present the 5-fold cross-validated Precision, Recall, and Mean Acc. under subject-based selecting scheme, resulting from each 3 classifiers (AlexNet, DenseNet, SE_Inception_v4) trained by DDMs generated from SampEn, FuzzyGMEn and FuzzyLMEn respectively. The method that reports the best score is

SE_Inception_v4 trained by FuzzyGMEn-generated DDMs, resulting in Acc. = 81.85% and Std. = 2.97%.

TABLE III

THE PERFORMANCE SUMMARY USING SAMPEN-GENERATED DDMS BASED ON SUBJECT-BASED SCHEME

DDM generation method	CNN model	Fold #	Precision (%)	Recall (%)	Acc. (%)	Mean Acc. ± Std. (%)
SampEn	*AlexNet*	fold1	72.95	53.97	76.99	*79.81±3.90*
		fold2	89.04	71.23	85.32	
		fold3	81.35	41.88	76.90	
		fold4	70.93	50.61	76.08	
		fold5	79.29	59.87	83.75	
SampEn	*DenseNet*	fold1	67.92	78.35	79.56	*78.43±3.04*
		fold2	65.14	86.31	76.59	
		fold3	75.48	41.33	75.41	
		fold4	78.45	43.86	76.75	
		fold5	77.42	71.70	83.86	
SampEn	*SE_Inception_v4*	fold1	76.24	54.25	78.17	*80.94±4.10*
		fold2	86.67	78.89	87.01	
		fold3	77.09	50.08	77.90	
		fold4	75.23	48.23	76.93	
		fold5	75.69	70.24	84.70	

TABLE IV

THE PERFORMANCE SUMMARY USING FUZZYGMEN-GENERATED DDMS BASED ON SUBJECT-BASED SCHEME

DDM generation method	CNN model	Fold #	Precision (%)	Recall (%)	Acc. (%)	Mean Acc. ± Std. (%)
Fuzzy GMEn	*AlexNet*	fold1	80.03	49.88	78.20	*80.09±2.94*
		fold2	86.96	76.06	85.52	
		fold3	85.80	44.95	78.69	
		fold4	76.06	48.76	77.29	
		fold5	78.09	57.41	80.75	
Fuzzy GMEn	*DenseNet*	fold1	81.21	50.54	78.69	*77.15±2.02*
		fold2	64.99	77.75	74.92	
		fold3	77.51	55.81	79.41	
		fold4	77.91	50.36	78.20	
		fold5	72.87	35.56	74.52	
Fuzzy GMEn	*SE_Inception_v4*	fold1	72.32	67.26	79.62	*81.85±2.97*
		fold2	86.73	79.02	87.07	
		fold3	77.05	57.72	79.72	
		fold4	81.56	51.77	79.56	
		fold5	86.46	49.74	83.27	

TABLE V

THE PERFORMANCE SUMMARY BY FUZZYLMEN-GENERATED DDMS BASED ON SUBJECT-BASED SCHEME

DDM generation method	CNN model	Fold #	Precision (%)	Recall (%)	Acc. (%)	Mean Acc. ± Std. (%)
FuzzyLMEn	*AlexNet*	fold1	77.13	48.47	77.04	*77.83±3.30*
		fold2	83.20	73.07	83.70	
		fold3	74.93	42.99	75.65	
		fold4	67.89	52.06	74.11	
		fold5	76.03	50.91	78.63	
FuzzyLMEn	*DenseNet*	fold1	78.18	52.51	78.35	*74.22±3.41*
		fold2	81.85	26.80	69.06	
		fold3	72.35	43.94	75.15	
		fold4	76.37	46.62	76.88	
		fold5	75.84	19.86	71.64	
FuzzyLMEn	*SE_Inception_v4*	fold1	80.50	48.47	77.95	*79.46±1.98*
		fold2	84.43	69.80	83.16	
		fold3	80.94	45.05	77.64	
		fold4	83.05	47.84	78.88	
		fold5	70.94	64.50	79.69	

B. SEGMENT-BASED SCHEME

Tables 6-8 present the results under segment-based selecting scheme. The method that reports the best score is SE_Inception_v4, which trained by global type data, resulting in Mean Acc. = 80.94% and Std. = 1.71%. Mean accuracies of all 3 trained models score between 78.05% and 80.94%, except for the lowest score of 76.82% generated by SampEn-generated DDMs. It is also shown that the performance for these three models increase greatly when the percent of data to train varies from 10% to 90%.

It is clear that inception-v4 performs the best with the highest mean accuracy for each of the 3 methods and both selecting schemes. It can also be seen that FuzzyGMEn-generated matrixes tend to show a more profound feature vector for distinguishing CHF and NSR subjects, which are classified with a higher accuracy compared with those of FuzzyLMEn-generated DDMs in Tables 5, 8 and SampEn-generated DDMs in Tables 3, 6, respectively.

TABLE VI

THE PERFORMANCE SUMMARY USING SAMPEN-GENERATED DDMS BASED ON SEGMENT-BASED SCHEME

DDM generation method	CNN model	Training Data (%)	Test Data (%)	Precision (%)	Recall (%)	Acc. (%)	Mean Acc. ± Std. (%)
SampEn	*AlexNet*	10	90	82.35	46.82	77.82	*78.05±0.85*
		20	80	82.27	45.77	77.51	
		30	70	78.42	53.29	78.46	
		40	60	78.17	54.88	78.80	
		50	50	78.10	55.19	78.85	
		60	40	79.59	52.30	78.54	
		70	30	80.51	51.27	78.56	
		80	20	77.67	52.07	77.93	
		90	10	69.35	46.32	75.97	
SampEn	*DenseNet*	10	90	74.47	48.12	76.01	*78.35±1.80*
		20	80	81.10	41.25	76.01	
		30	70	74.68	63.28	79.59	
		40	60	83.11	45.29	77.58	
		50	50	82.24	41.75	76.40	
		60	40	80.42	55.71	79.68	
		70	30	82.52	56.35	80.51	
		80	20	80.17	60.13	80.78	
		90	10	78.12	53.97	78.55	
SampEn	*SE_Inception_v4*	10	90	70.08	61.75	77.33	*79.28±0.97*
		20	80	88.49	44.69	78.55	
		30	70	76.21	62.43	79.98	
		40	60	82.05	55.64	80.17	
		50	50	85.20	51.58	79.87	
		60	40	80.83	55.25	79.68	
		70	30	78.02	60.18	80.09	

80	20	85.03	51.42	79.76	
90	10	83.22	46.99	78.08	

TABLE VII

THE PERFORMANCE SUMMARY USING FUZZYGMEN-GENERATED DDMS

BASED ON SEGMENT-BASED SCHEME

DDM generation method	CNN model	Training Data (%)	Test Data (%)	Precision (%)	Recall (%)	Acc. (%)	Mean Acc. ± Std. (%)
Fuzzy GMEn	*AlexNet*	10	90	84.04	45.67	77.90	*79.27± 1.04*
		20	80	75.33	55.51	78.01	
		30	70	84.00	52.86	79.92	
		40	60	82.87	51.11	79.13	
		50	50	75.40	54.61	77.83	
		60	40	76.02	62.94	80.02	
		70	30	81.96	55.55	80.13	
		80	20	82.42	57.34	80.73	
		90	10	79.97	56.56	79.79	
Fuzzy GMEn	*DenseNet*	10	90	84.91	32.17	74.37	*79.63± 2.69*
		20	80	78.85	42.73	75.88	
		30	70	79.04	62.79	81.10	
		40	60	80.17	55.14	79.48	
		50	50	79.17	54.27	78.95	
		60	40	84.27	54.52	80.46	
		70	30	79.71	68.74	82.90	
		80	20	78.14	68.32	82.16	
		90	10	80.99	61.15	81.34	
Fuzzy GMEn	*SE_Inception_v4*	10	90	72.60	54.54	76.83	*80.94± 1.71*
		20	80	78.28	60.29	80.19	
		30	70	80.66	60.65	81.09	
		40	60	82.98	55.87	80.50	
		50	50	86.34	56.01	81.46	
		60	40	75.77	68.88	81.34	
		70	30	82.18	66.74	83.26	
		80	20	85.51	60.81	82.62	
		90	10	79.39	62.68	81.20	

TABLE VIII

THE PERFORMANCE SUMMARY USING FUZZYLMEN-GENERATED DDMS

BASED ON SEGMENT-BASED SCHEME

DDM generation method	CNN model	Training Data (%)	Test Data (%)	Precision (%)	Recall (%)	Acc. (%)	Mean Acc. ± Std. (%)
FuzzyLMEn	AlexNet	10	90	75.02	39.37	74.13	*76.82± 1.30*
		20	80	74.87	50.85	76.76	
		30	70	81.30	47.18	77.66	
		40	60	74.73	55.59	77.83	
		50	50	81.97	46.62	77.68	
		60	40	72.88	58.02	77.68	
		70	30	75.19	53.52	77.51	
		80	20	71.01	59.80	77.31	
		90	10	73.08	44.69	74.82	
FuzzyLMEn	DenseNet	10	90	73.48	46.76	75.40	*78.16± 1.56*
		20	80	80.34	42.19	76.09	
		30	70	89.68	46.22	79.26	
		40	60	77.72	52.92	78.16	
		50	50	78.68	48.13	77.23	
		60	40	83.25	48.08	78.37	
		70	30	77.10	61.82	80.17	
		80	20	76.15	62.88	80.05	
		90	10	74.64	59.43	78.68	
FuzzyLMEn		10	90	74.06	49.06	76.10	*79.73± 1.59*
		20	80	72.05	62.85	78.41	
		30	70	74.83	63.89	79.79	
		40	60	81.78	56.43	80.31	
		50	50	82.59	58.73	81.18	
		60	40	75.06	65.50	80.24	
		70	30	80.50	62.40	81.52	
		80	20	74.57	69.48	80.96	
		90	10	79.72	54.16	79.09	

V. DISCUSSION

In this study, we choose three CNN models for classifying the NSR and CHF patients, and compared their performances. The result shows that no matter what models we choose, the performances of three model have no significant difference. This means the result is not an accidental phenomenon based on one model. We also choose two different schemes to train models. Under the subject-based scheme, training and test data are totally independent. Under segment-based scheme, a certain fraction of each subject's segments is randomly selected as the training set and the remaining are used as the test set. Previous study has proved models trained by dependent data performed much better than models trained by independent data [28]. However, in this study, the results from the segment-based scheme are similar to the results from the subject-based scheme. This is due to the large intra-subject variability of DDMs.

Over the past years, automatic classifiers have been proposed in diagnosing patients who are suffering CHF. Isler et al. proposed a model based on KNN and wavelet entropy measures of HRV indices [12]. When they used all features to train models, their accuracy is between 78.31% and 84.34%. However, after they used genetic algorithm (GA) for feature selection, they obtained an accuracy as high as 96.39%. However, the method is too complicated for the daily monitoring. A classifier based on classification and regression tree (CART) was proposed by Pecchia et al. to distinguish CHF patients from NSR subjects. This method is simpler and can be fully understood without advanced mathematical skills. They evaluate the result of CART to choose feature and discriminate CHF patients. It is worth mentioning that they use "tree A" to classify segments and then use "tree B" to classify subjects. Therefore, their final result is to evaluate the performance of classifying subjects.

The difference between our study and other studies is that, we trained the model for CHF segments classification, not for CHF patients classification. In this way, our performance result cannot be compared with their result because we are measuring different things. This research also allowed us a further research direction: seek for the proper ratio of abnormal segments for CHF diagnosis. Jovic et al. proposed a model based on random forest and combinations of linear and nonlinear features of HRV [13]. They achieved an accuracy of 73% when they only used combinations of entropy calculation result as the input of the classifier. This result can be improved to around 84% by using combinations of linear and non-linear HRV features. This unpromising result by simply using the combinations of entropy calculation can also prove that DDM contains more information than simple entropy calculation. There are also researchers who designed classifiers based on SVM method and combination of several HRV features and reached high accuracy [14-16]. Liu et al. [15] and Wang et al. [14] compared the contributions of different combinations of HRV features to performance of classifiers. Liu et al. reached a

highest accuracy of 91.49% using combination of time domain and non-linear features, which is consistent with the conclusion of Jovic et al. [13].

All these studies are using multiple features as the input of classifiers, for the reason that the performance with single feature is far poorer. Jovic et al. [13] achieved results between 60% and 75% which are far lower than other results by using combination of the same type of features, such as approximate entropy (ApEn1-ApEn4), maximum approximate entropy (MaxApEn), multiscale sample entropy (SampEn1-SampEn20), multiscale carnap 1D entropy(Carnap1-Carnap20). The above features all belong to the entropy method category but their calculation methods are different. The result of previous studies depends on which feature set is chosen. However, this best choice may change when choosing different datasets. Additionally, it is also too complex and demanding for the daily activity of clinicians.

VI. CONCLUSION

In our study, we only used one feature to train models and obtained the highest accuracy of 81.85%. This result is much higher compared to the result of using combination of the same type of features. However, it is much lower than the previous studies which are using combinations of different features. For the next step, we plan to add other dimension images to improve the completeness of input and we expect the result will be improved. Single dimension of input is still too 'thin or lean' for a model to train, which can be seen in the current result. Adding more dimension images does not mean we will increase steps of feature selection, since it is CNN itself that extract features. We can also train the CNN classifier using larger dataset, for the reason that small datasets will cause the deep neural network to overfit.

APPENDIX

A. SAMPLE ENTROPY METHOD (SAMPEN)

For RR segment $x(i)$ $(1 \le i \le N)$, form the vector sequences X_i^m:

$X_i^m = \{x(i), x(i+1), \ldots, x(i+m-1)\}, 1 \le i \le N-m+1$

Then the distance between X_i^m and X_j^m based on the maximum absolute difference is defined as:

$$d_{i,j}^m = d\left[X_i^m, X_j^m\right] = \max_{k=0}^{m-1} \left| x(i+k) - x(j+k) \right|$$

In SampEn, if the distance is within the threshold parameter r = 0.2, the similarity degree between the two vectors is 1; if the distance is beyond the threshold parameter r, the similarity degree is 0. There is absolutely a 0 or 1 determination.

B. FUZZY MEASURE ENTROPY (FUZZYMEN)

For RR segment $x(i)$ $(1 \le i \le N)$, firstly form the local vector sequences XL_i^m and global vector sequences XG_i^m respectively:

$$XL_i^m = \{x(i), x(i+1), \cdots, x(i+m-1)\} - \bar{x}(i)$$
$$XG_i^m = \{x(i), x(i+1), \cdots, x(i+m-1)\} - \bar{x}$$
$$1 \le i \le N-m$$

The vector XL_i^m represents m consecutive $x(i)$ values but removing the local baseline $\bar{x}(i)$, which is defined as:

$$\bar{x}(i) = \frac{1}{m}\sum_{k=0}^{m-1} x(i+k) \quad 1 \le i \le N-m$$

The vector XG_i^m also represents m consecutive $x(i)$ values but removing the global mean value $\bar{x}$ of the segment $x(i)$, which is defined as:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x(i)$$

Then the distance between the local vector sequences XL_i^m and XL_j^m and the distance between the global vector sequences XG_i^m and XG_j^m are defined as follows respectively:

$$dL_{i,j}^m = d\left[XL_i^m, XL_j^m\right] = \max_{k=0}^{m-1} \left| (x(i+k) - \bar{x}(i)) - (x(j+k) - \bar{x}(j)) \right|$$
$$dG_{i,j}^m = d\left[XG_i^m, XG_j^m\right] = \max_{k=0}^{m-1} \left| (x(i+k) - \bar{x}) - (x(j+k) - \bar{x}) \right|$$

Given the parameters nL, rL, nG and rG, calculate the similarity degree $DL_{i,j}^m(n_L, r_L)$ between the local vectors XL_i^m and XL_j^m by the fuzzy function $\mu L(dL_{i,j}^m, n_L, r_L)$, as well as calculate the similarity degree $DG_{i,j}^m(n_G, r_G)$ between the global vectors XG_i^m and XG_j^m by the fuzzy function $\mu G(dG_{i,j}^m, n_G, r_G)$:

$$DL_{i,j}^m(n_L, r_L) = \mu L(dL_{i,j}^m, n_L, r_L) = \exp\left(-\frac{(dL_{i,j}^m)^{n_L}}{r_L}\right)$$
$$DG_{i,j}^m(n_G, r_G) = \mu G(dG_{i,j}^m, n_G, r_G) = \exp\left(-\frac{(dG_{i,j}^m)^{n_G}}{r_G}\right)$$

In this study, the local similarity weight n_L=1 and global vector similarity weight n_G=2, the local tolerance threshold r_L was set equal to the global threshold r_G, i.e., $r_L = r_G = r$.

REFERENCE

[1] J. K. Ghali, R. Cooper, and E. Ford, "Trends in hospitalization rates for heart failure in the United States, 1973–1986," *Archives of Internal Medicine*, vol. 150, no. 4, pp. 769-773, 1990.

[2] R. F. Gillum, "Heart failure in the United States 1970–1985," *American Heart Journal*, vol. 113, no. 4, pp. 1043-1045, 1987.

[3] A. Mosterd and A. W. Hoes, "Clinical epidemiology of heart failure," *Heart*, vol. 93, no. 9, pp. 1137-1146, 2007.

[4] D. D. Schocken, M. I. Arrieta, P. E. Leaverton, and E. A. Ross, "Prevalence and mortality rate of congestive heart failure in the United States," *Journal of the American College of Cardiology*, vol. 20, no. 2, pp. 301-306, 1992.

[5] F. Hobbs, J. Doust, J. Mant, and M. R. Cowie, "Diagnosis of Heart Failure in Primary Care," *Heart*, vol. 96, no. 21, pp. 1773-1777, 2010.

[6] R. Cardarelli and T. G. Lumicao, "B-type natriuretic peptide: a review of its diagnostic, prognostic, and therapeutic monitoring value in heart failure for primary care physicians," *The Journal of the American Board of Family Practice*, vol. 16, no. 4, pp. 327-333, 2003.

[7] A. Fuat, A. P. S. Hungin, and J. J. Murphy, "Barriers to accurate diagnosis and effective management of heart failure in primary care: qualitative study," *British Medical Journal*, vol. 326, no. 7382, p. 196, 2003.

[8] J. Remes, H. MlEttinen, A. Reunanen, and K. Pyörälä, "Validity of clinical diagnosis of heart failure in primary health care," *European Heart Journal*, vol. 12, no. 3, pp. 315-321, 1991.

[9] J. Nolan, P. D. Batin, R. Andrews, S. J. Lindsay, P. Brooksby, M. Mullen, W. Baig, A. D. Flapan, A. Cowley, and R. J. Prescott, "Prospective study of heart rate variability and mortality in chronic heart failure: results of the United Kingdom heart failure evaluation and assessment of risk trial (UK-heart)," *Circulation Journal*, vol. 98, no. 15, pp. 1510-1516, 1998.

[10] M. Hadase, A. Azuma, K. Zen, S. Asada, T. Kawasaki, T. Kamitani, S. Kawasaki, H. Sugihara, and H. Matsubara, "Very low frequency power of heart rate variability is a powerful predictor of clinical prognosis in patients with congestive heart failure," *Circulation Journal*, vol. 68, no. 4, pp. 343-347, 2004.

[11] M. T. La Rovere, G. D. Pinna, R. Maestri, A. Mortara, S. Capomolla, O. Febo, R. Ferrari, M. Franchini, M. Gnemmi, and C. Opasich, "Short-term heart rate variability strongly predicts sudden cardiac death in chronic heart failure patients," *Circulation Journal*, vol. 107, no. 4, pp. 565-570, 2003.

[12] Y. Isler and M. Kuntalp, "Combining classical HRV indices with wavelet entropy measures improves to performance in diagnosing congestive heart failure," *Computers in Biology and Medicine*, vol. 37, no. 10, pp. 1502-1510, 2007.

[13] A. Jovic and N. Bogunovic, "Random forest-based classification of heart rate variability signals by using combinations of linear and nonlinear features," in *XII Mediterranean Conference on Medical and Biological Engineering and Computing 2010*, pp. 29-32: Springer, 2010.

[14] Y. Wang, S. Wei, S. Zhang, Y. Zhang, L. Zhao, C. Liu, and A. Murray, "Comparison of time-domain, frequency-domain and non-linear analysis for distinguishing congestive heart failure patients from normal sinus rhythm subjects," *Biomedical Signal Processing and Control*, vol. 42, pp. 30-36, 2018.

[15] G. Liu, L. Wang, Q. Wang, G. Zhou, Y. Wang, and Q. Jiang, "A new approach to detect congestive heart failure using short-term heart rate variability measures," *PloS One*, vol. 9, no. 4, p. e93399, 2014.

[16] G. Yang, Y. Ren, Q. Pan, G. Ning, S. Gong, G. Cai, Z. Zhang, L. Li, and J. Yan, "A heart failure diagnosis model based on support vector machine," in *Biomedical Engineering and Informatics (BMEI), 3rd International Conference on*, vol. 3, pp. 1105-1108: IEEE, 2010.

[17] L. Pecchia, P. Melillo, M. Sansone, and M. Bracale, "Discrimination power of short-term heart rate variability measures for CHF assessment," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 1, pp. 40-46, 2011.

[18] S. M. Pincus and A. L. Goldberger, "Physiological time-series analysis: what does regularity quantify?," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 266, no. 4, pp. H1643-H1656, 1994.

[19] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039-H2049, 2000.

[20] R. Barbieri, E. P. Scilingo, and G. Valenza, *Complexity and Nonlinearity in Cardiovascular Signals*. Springer, 2017.

[21] L. Zhao, S. Wei, C. Zhang, Y. Zhang, X. Jiang, F. Liu, and C. Liu, "Determination of Sample Entropy and Fuzzy Measure Entropy Parameters for Distinguishing Congestive Heart Failure from Normal Sinus Rhythm Subjects," *Entropy*, vol. 17, no. 12, pp. 6270-6288, 2015.

[22] C. Liu and R. Gao, "Multiscale entropy analysis of the differential RR interval time series signal and its application in detecting congestive heart failure," *Entropy*, vol. 19, no. 6, p. 251, 2017.

[23] A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation Journal*, vol. 101, no. 23, pp. 215-220, 2000.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097-1105, 2012.

[25] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, p. 3, 2017.

[26] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *the Association for the Advance of Artificial Intelligence (AAAI)*, vol. 4, p. 12, 2017.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *European Conference on Computer Vision*, 2015.

[28] Q. Qin, J. Li, L. Zhang, Y. Yue, and C. Liu, "Combining Low-dimensional Wavelet Features and Support Vector Machine for Arrhythmia Beat Classification," *Scientific Reports*, vol. 7, no. 1, p. 6067, 2017.

[29] D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman, "Sample entropy analysis of neonatal heart rate variability," *American Journal of Physiology-Regulatory Integrative and Comparative Physiology*, vol. 283, no. 3, pp. R789-797, 2002.

[30] W. T. Chen, J. Zhuang, W. X. Yu, and Z. Z. Wang, "Measuring complexity using FuzzyEn, ApEn, and SampEn," *Medical Engineering and Physics*, vol. 31, no. 1, pp. 61-68, 2009.

[31] C. Y. Liu, K. Li, L. N. Zhao, F. Liu, D. C. Zheng, C. C. Liu, and S. T. Liu, "Analysis of heart rate variability using fuzzy measure entropy," *Computers in Biology and Medicine*, vol. 43, no. 2, pp. 100-108, 2013.

[32] M. U. Ahmed and D. P. Mandic, "Multivariate multiscale entropy: a tool for complexity analysis of multichannel data," (in eng), *Physical Review E*, vol. 84, no. 6 Pt 1, p. 061918, 2011.

[33] C. Y. Liu and L. N. Zhao, "Using Fuzzy Measure Entropy to improve the stability of traditional entropy measures," in *Computing in Cardiology*, Hangzhou, vol. 31, pp. 681 - 684, 2011.

[34] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, and M. Devin, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2015, Software available from tensorflow.org .

[35] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," 2011.