Multi-model Deep Learning Ensemble for ECG Heartbeat Arrhythmia Classification

Ehab Essa[1,2], Xianghua Xie[1]
[1]*Department of Computer Science*, *Swansea University*, Swansea, UK
[2]*Department of Computer Science, Faculty of Computers and Information, Mansoura University*, Mansoura, Egypt
Email: e.m.m.essa@swansea.ac.uk, x.xie@swansea.ac.uk

***Abstract*—Managing and treating cardiovascular diseases can be substantially improved by automatic detection and classification of the heart arrhythmia. In this paper, we introduced a novel deep learning system for classifying the electrocardiogram (ECG) signals. The heartbeats are classified into different arrhythmia types using two proposed deep learning models. The first model is integrating the convolutional neural network (CNN) and long short-term memory (LSTM) network to extract useful features within the ECG signal. The second model combines several classical features with LSTM in order to effectively recognize abnormal classes. These deep learning models are trained using a bagging model then aggregated by a fusion classifier to form a robust unified model. The proposed system is evaluated on the MIT-BIH arrhythmia database and produces an overall accuracy of 95.81%, which significantly outperforms the state-of-the-art.**

***Index Terms*—CNN, LSTM, Bagging, Deep Learning Ensemble, ECG, Arrhythmia**

I. Introduction

Electrocardiography (ECG) is a non-invasive diagnostic technique for arrhythmias and conduction disorders by measuring the heart rhythm and its electrical activity over a period of time. ECG detects the electrical signals generated by heartbeats using a set of sensors called electrodes that are attached to the skin. Typically 12-lead configuration [1] is used to record ECG with 10 electrodes distributed on the patient's limb and chest. In the case of a long-term continuous heart monitoring, 2-lead ECG is commonly used with a Holter monitoring device. ECG waveform is examined beat-by-beat by a trained specialist to detect arrhythmias. This can be a very time-consuming and tedious process, especially each ECG recording may take several minutes to several hours. Therefore, developing an automated system for analyzing and diagnosing heart arrhythmias is highly desirable.

In this paper, we propose a cascade deep learning method based on convolutional neural networks (CNN) and long short-term memory (LSTM) for automatic heartbeat arrhythmia classification. At the first stage, we employ a bag of LSTM models that use both CNN for feature extraction and classical descriptive features, i.e. the RR intervals and higher-order statistics (HOS). The two sets of features are introduced separately to the LSTM bagging models. In the second stage, all the LSTM based bagging models are combined into a new deep neural network to fuse the classification results of the different LSTM models. In the last stage, another CNN-LSTM model is proposed to reduce the false positives produced by the previous stage. The method is evaluated on the MIT-BIH arrhythmia database and follows the Association for the Advancement of Medical Instrumentation (AAMI) [2] recommendations. This arrhythmia database is highly imbalanced, which poses a challenge to recognize minority classes. We tackle this challenge in two ways: the use of bagging and our weighted loss function. The bagging model is introduced to alter the training distributions in order to tackle the level of imbalance. The weighted loss function gives more weights for minority classes to encourage the classifier to recognize them correctly.

II. Related works

Various techniques have been developed to classify ECG heartbeats in the literature. Most of these techniques include the following steps: pre-processing, feature extraction and classification. Various classification techniques have been used to label the extracted ECG features, such as support vector machine (SVM) [3]–[6], decision tree [7], linear discriminants (LDs) [8], and deep learning methods [9]–[11]. For example, in [8], Chazal et al. used linear discriminants to classify heartbeats into five classes. The method investigates a different combination of features based on ECG morphology, heartbeat intervals, and RR intervals. In [3], Ye et al. utilized independent component analysis and wavelet transform to extract morphological features and combined them with RR intervals features to classify heartbeats into 16 classes. The features are extracted from two ECG leads and used to train different SVM classifiers independently, then the final classification result was obtained by fusing the decisions of SVM classifiers. In [5], Raj et al. proposed to extract the ECG features by using the sparse representation technique with a Gabor dictionary. A set of features are computed from each of the significant atoms of the dictionary and concatenated to form a feature vector. The authors used the least-square twin SVM model to classify the extracted features where the particle swarm optimization (PSO) was adopted to optimize the learning parameters. However, the performance of these methods is still limited.

Ensemble-based methods have been proposed in order to improve the overall performance of the classification problem. Zhang et al. [4] used an ensemble of SVMs to automatically detect heart arrhythmia. Features such as ECG morphology and interval characterization are extracted from two leads and

trained separately on several one-vs-one SVM models. Then the final decision is reached by using the product rule to combine the decisions of different models. Mondéjar-Guerra et al. [6] extended the previous method by creating a dedicated SVM classifier for each type of features and testing a different set of features such as local binary patterns (LBP), HOS, wavelets, and RR intervals. Shi et al. [7] proposed a hierarchical classification method based on XGBoost classifiers. A large set of features were extracted, then the method employed a recursive feature elimination to select a subset of features used to train the classifier. However, these methods are heavily depending on the use of many hand-crafted features that may not be well generalized.

Deep learning methods, e.g. CNN, have received a lot of attention in recent years because of its outstanding performance in fields such as computer vision and medical image analysis [12], [13] compared to traditional machine learning approaches [14], [15]. Sellami et al. [9] presented a CNN method with a batch-weight loss to reduce the effect of imbalance between classes. The method was trained on the raw ECG signals by taking two consecutive heartbeats; the target heartbeat was proceeded by its prior heartbeat. However, the accuracy of the majority class is highly decreased and the false positive rate is still high. Jiang et al. [16] introduced multi-module deep learning method to handle imbalance data by over-sampling the minority classes and using features generated by an auto-encoder to train a CNN to classify the heartbeats into four classes. However, the method becomes more complex and very slow due to the over-sampling strategy, which may also lead to overfitting. Mathews et al. [11] used restricted Boltzmann machines (RBM) and deep belief networks (DBN) to automatic classify ECG signals. The RBM is initially trained on a set of hand-crafted ECG features. DBN is formed as a stack of RBMs where each RBM is considered as a hidden layer and learned from the output of the previous RBM model in the stack. However, the method is still mainly relying on the hand-crafted features which affect its overall accuracy for recognizing patient-independent ECG data. Xu et al. [10] proposed a similar deep learning approach, but instead of initially modeling a set of hand-crafted ECG features, the DBN works on the raw ECG signal. However, the method is not exploiting the long-term time-dependencies of ECG data to extract more meaningful features.

III. Proposed method

First, the ECG signals are pre-processed and segmented into heartbeats. Each heartbeat class is learned by a set of binary classifiers using the one-vs-all scheme. The heartbeats are fed into CNN-LSTM and RRHOS-LSTM bagging models belonging to the first heartbeat class to get the probabilities that the beats belong to the given class. The output of the bagging models serves as an input to the fusion classifier to lead the final decision. If the beat is classified as positive, then it moves to the verification network to confirm its label. If it is classified as negative by the fusion classifier or the verification network, then it moves to the next bagging models for the second class and so on.

A. Pre-processing

In the literature [4], [6], [8], there are two common pre-processing steps: baseline removal and high-frequency noise reduction. Since the MIT-BIH database contains ECG signals acquired by Holter devices, the signals were susceptible to baseline wandering and high-frequency noise. The baseline of ECG is computed by applying two median filters one after another of size 200-ms and 600-ms. The baseline is then subtracted from the original ECG signal to create the baseline-corrected signal. The second pre-processing step preforms the high-frequency and power-line noise removal by applying 12-order low-pass filter with cutoff frequency at 35 Hz. The filtered ECG signals were used in all further processing. The ECG signal comprises a sequence of heartbeats. In this paper, the annotations of QRS complex that comes with the MIT-BIH database were utilized to obtain the heartbeats. Here, we take a window of 180 samples for each heartbeat, where R peak is located at its center.

B. Bagging deep learning models

In this paper, we propose to build and train an ensemble of two deep learning models for the classification of the heartbeat signal. CNN-LSTM is the first deep learning model of the proposed method, based on a combination of CNN and LSTM. The second model is RRHOS-LSTM which integrates the RR intervals and HOS features with LSTM. Bagging of CNN-LSTM and RRHOS-LSTM is introduced to create an ensemble model to improve the model robustness and tackle the data imbalance problem.

The MIT-BIH database is a highly unbalanced dataset. Here, we apply randomness in bagging to increase generalization capability. First, we convert the multi-class classification problem to binary classification using the one-vs-all scheme. Then, for each binary class model, the negative data is randomly down-sampled while keeping the positive data the same to create a new training dataset. The process is repeated multiple times to create many sub-sampling training sets. These datasets are used to train two deep learning models to get a set of different models not only in the architecture but also in the data-level. The weighted loss function is also employed in each deep learning model to prevent the model from learning only the majority class.

1) CNN-LSTM model: CNN has several useful properties, such as local connectivity, weight sharing, and spatial pooling, which significantly decreases the number of parameters compared to the fully connected network and helps to learn local translational invariant features. LSTM is considered one of the main preferred neural network architecture for modeling time series data. LSTM differs from other deep neural networks in that neural output from the current time step is connected to the inputs of the next time step. This enables the LSTM to maintain the internal state to process sequential input.

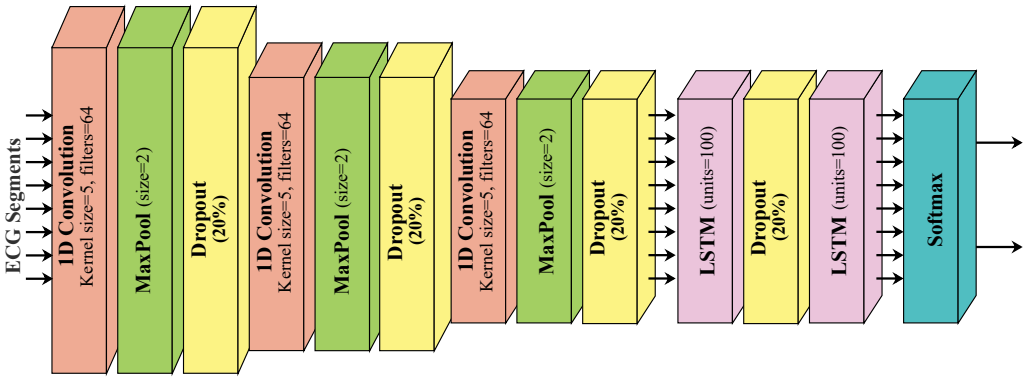

Fig. 1. The proposed CNN-LSTM architecture.

Here, CNN and LSTM are proposed to learn high-level hierarchical features from the ECG signal. The CNN-LSTM network consists of 3 convolutional layer, 3 max-pooling, and 2 LSTM layers, as shown in Figure 1. The network takes 5 consecutive heartbeats as an input, where the current heartbeat is located at the center. The input size of the CNN-LSTM network is 900 samples. Each convolutional layer has a kernel size of 5 with stride of 1. The kernel is moving over the input sequence one step (i.e. stride equals to 1) at a time and convoluted with the corresponding elements of the input. The number of output features maps (i.e. filters) of each convolutional layer is 64. The ReLU activation function is used for all three convolutional layers. Max-pooling is applied after every convolutional layer with pooling size is 2 and stride is 2. This reduces the feature map size by half. After each max-pooling layer, we add a dropout of 20% to reduce the network over-fitting.

The final structure of the CNN-LSTM network is 2 LSTM layers, followed by a softmax layer to predict the output class. The temporal information is extracted by the LSTM layers from CNN feature maps. Each LSTM unit has 100 hidden neurons. The output of the first LSTM layer is a sequence of the hidden units corresponding to each input time step. The second LSTM layer returns the hidden units of the last time step. The output of the last LSTM layer is fed to a softmax function to produce the probabilities of each output class.

2) RRHOS-LSTM model: RR intervals are the most common classical features used to characterize ECG signals according to the literature [3], [4], [6]–[8], [11]. RR intervals refer to the time between R peak points of consecutive heartbeats. There are four RR intervals: pre-RR, post-RR, local-RR, and average-RR. HOS refers to kurtosis and skewness, which measures asymmetry and sharpness of a given heartbeat. In this paper, each heartbeat is segmented into 6 intervals, and then the kurtosis and skewness are computed for each one.

The RRHOS-LSTM model combines the classical features (i.e. RR intervals and HOS) with LSTM to classify heartbeats arrhythmia. The RRHOS-LSTM composes of a feature extraction layer and one LSTM layer. RR intervals are extracted to highlight local and global information about R-peak of two consecutive heartbeats. RR intervals have 8 features that are computed from the given heartbeat. HOS represents high-order statistical information of the heartbeat. HOS is computed from five consecutive heartbeats where the current heartbeat is surrounded by four heartbeats that are equally distributed on both sides. The number of HOS features is 60. Therefore the total number of features fed to LSTM as input is 68. The LSTM is processing the input features to learn temporal dependencies. The LSTM layer produces the hidden states output for each time step, which is then flattened and fed to a softmax layer to classify the heartbeat signal.

C. Classifier Fusion

The fused classifier is a meta-learner that collects the output of the bagging models to form an ensemble model. This classifier is a deep neural network that takes its input from the probability output of heartbeat classification generated by CNN-LSTM and RRHOS-LSTM bagging models. Suppose we have N bagging models of CNN-LSTM and RRHOS-LSTM classifier, thereby the length of the input vector of the fusion classifier is $2 * N$ where each bagging model has two outputs.

The fusion classifier consists of a batch normalization layer, two fully connected layers, and a softmax layer. Each fully connected layer has 500 hidden neurons and used ReLU as an activation function. Dropout layer with a ratio of 20% is applied between the two fully connected layers. In the last layer, a softmax function is employed to produce the arrhythmia classification of the heartbeat. The number of output neurons is 2, one of the positive class and the other for the rest classes, as the fusion classifier is trained using the one-vs-all scheme.

D. Verification network

At last stage, we propose a verification deep neural network to validate the output of the fusion classier in order to minimize false positive. Verification network is based on CNN and LSTM, similar to the CNN-LSTM network. Verification network consists of 3 convolutional layers, where each one is followed by max-pooling layer and two LSTM layer. The

TABLE I
THE PERFORMANCE OF CNN-LSTM AND RRHOS-LSTM BAGGING MODELS INDIVIDUALLY AND AFTER COMBINING THEM USING THE FUSION CLASSIFIER AND THEN REFINING THE FUSION MODEL USING THE VERIFICATION NETWORK.

	SVEB					VEB					F				
	Se	Sp	PPv	F1	Acc	Se	Sp	PPv	F1	Acc	Se	Sp	PPv	F1	Acc
CNN-LSTM	81.63	84.43	18.41	30.04	84.31	94.50	98.74	83.88	88.87	98.47	45.10	93.35	5.08	9.13	92.97
RRHOS-LSTM	84.27	68.58	10.35	18.43	69.22	94.62	96.32	64.08	76.41	96.21	86.60	84.51	4.22	4.22	84.52
Majority Voting	83.44	78.44	14.28	24.39	78.65	94.93	99.53	93.31	94.11	99.23	83.51	91.46	7.15	13.18	91.39
Fusion Classifier	80.12	86.67	20.55	32.71	86.40	95.62	99.38	91.41	93.47	99.13	38.40	98.18	14.26	20.80	97.71
Verification Net.	65.51	98.56	66.19	65.85	97.20	93.91	99.62	94.55	94.23	99.25	19.33	99.79	41.67	26.41	99.16

difference between the verification network and the previously mentioned CNN-LSTM is the dropout regularization. In the verification network, we add two dropout layers, one between the last layer of the CNN part and first LSTM layer and the other between the last LSTM layer and the softmax layer. The rate of dropout sets to 50%. This order helps to reduce the over-fitting of the network. We finally combine all the proposed models in a cascade system to reach the final classification decision.

IV. RESULTS AND DISCUSSION

A. MIT-BIH arrhythmia Database

The MIT-BIH database [17], [18] is widely used as a standard ECG arrhythmia dataset for evaluating the performance of heartbeat classification. This dataset includes 48 two-lead ECG records of roughly 30 min acquired from 47 patients. Each record is sampled at 360 Hz and characterized by a set of labels marked at the R-peak of every beat. The modified-lead II (MLII) signals is used here. There are 16 heartbeat types in the original database. AAMI recommends grouping these heartbeats types into five categories: Normal (N), Supraventricular ectopic beat (SVEB), Ventricular ectopic beat (VEB), Fusion (F), and Unknown beat (Q). However, the Q class has a relatively too small number of samples (i.e. only 12 samples) especially after removing the 4 paced records. We choose to work on the other four AAMI classes and excluding Q class. In this paper, we use a subject-oriented patient independent evaluation scheme. We follow Chazal et al. [8] data division scheme to split the database to the training set (DS1) and testing set (DS2) to maintain inter-patient variation. Each dataset contains 22 ECG records from different patients with roughly the same ratio of beat types. This evaluation allows a fair comparison between different heartbeat classification methods.

B. Classification Performance

The performance of the proposed method is evaluated for each heartbeat category using 5 different metrics. These metrics are accuracy (Acc %), positive predictive value (PPv %), sensitivity (Se %), specificity (Sp %), and F1 score (F1 %). We used $1:4$ sampling ratio for training bagging classifiers where the number of negative samples is four times larger than the positive samples.

Firstly, a comparison between CNN-LSTM and RRHOS-LSTM bagging models is performed and reported in Table I (row 1 and 2). It shows CNN-LSTM bagging model is better than the RRHOS-LSTM bagging model in terms of the accuracy metric. However, the RRHOS-LSTM has better sensitivity for all classes that would boost the overall combined model.

Next, we compare between using the proposed fusion classifier and using the majority voting across the different classifiers as shown in Table I. The majority voting in Table I (row 3) combines the results of individual bagging models of CNN-LSTM and RRHOS-LSTM. Combining both bagging models shows a notable improvement, especially for the VEB and F classes. The result of the fusion classifier is reported in row 4, Table I. The fusion classifier has a significant improvement than the majority voting. The fusion classifier provides high levels of accuracy, 86.40% and 97.71% for SVEB and F classes, respectively compared to 78.65% and 91.39%. For the VEB class, the fusion classifier has a high accuracy of 99.13% that is very close to the accuracy of using the majority voting (99.23%). Overall, the accuracy performance of the fusion classifier is much higher than the individual CNN-LSTM and RRHOS-LSTM bagging models in all classes.

The following experiment is refining the result of the fusion classifier by using the proposed verification network, as shown in row 5, Table I. The verification network significantly improves the overall performance of the proposed system for all classes and reduces the false positive rate. For class SVEB, the positive predictive value and F1 score of the verification network (66.19% and 65.85%) are much higher than the fusion classifier (20.55% and 32.71%) with very high specificity 98.56% compared to 86.67% for the fusion classifier. For VEB class, the positive predictive value and F1 score after using the verification network is improved (94.55% and 94.23%) compared to without using it (91.41% and 93.47%). As a result of reducing the false positive, the sensitivity of the verification network is decreased for all classes compared to the fusion classifier.

The proposed method is also compared to 8 different state-of-the-art methods that follow the same evaluation scheme using patient independent on the MIT-BIH DS2 dataset. Table II illustrate the comparison between the proposed method and the following methods: ensemble SVM [6], Zhang et al [4], Shi et al [7], Raj & Ray [5], Sellami & Hwang [9], Mathews et al [11], Ye et al [3], and Chazal et al [8]. In Table II, the average of sensitivity, specificity, positive predictive value, and F1

TABLE II
THE AVERAGE VALUE OF THE EVALUATION METRICS OVER ALL HEARTBEAT TYPES AND THE OVERALL ACCURACY OF THE PROPOSED METHOD AND THE STATE-OF-THE-ART METHODS.

	Se	Sp	PPv	F1	Overall Acc.
The proposed method	69.20	94.56	74.97	71.06	95.81
Ensemble SVM [6]	70.29	95.55	66.35	67.09	94.47
Zhang et al [4]	86.82	95.42	60.36	64.02	88.34
Shi et al [7]	85.07	96.98	61.89	67.89	91.87
Raj & Ray [5]	87.50	96.47	61.50	65.69	90.27
Sellami & Hwang [9]	82.72	95.04	57.01	64.25	88.35
Mathews et al [11]	83.66	93.26	51.77	54.04	75.50
Ye et al [3]	62.79	92.61	53.46	55.96	86.55
Chazal et al [8]	83.19	95.13	56.98	60.12	86.24

score over all heartbeat types is reported and also the overall accuracy. The proposed method achieved the highest average positive predictive value and average F1 score (74.97% and 71.06%), the highest PPv for classes SVEB, VEB, and F, the highest specificity in SVEB, VEB, and F (98.56%, 99.62%, and 99.79%), and the highest sensitivity for class N (98.03%) and ranked the third for VEB class (93.91%). Overall, the proposed method achieved much better performance compared to all other methods, with an overall accuracy of 95.81%.

V. CONCLUSION

In this paper, an ensemble deep learning method is introduced for ECG heartbeat classification. The proposed system combines two different deep learning models. Both CNN and LSTM are used in the first model to extract dynamic features from the raw signal. The classical feature such as RR intervals and HOS are used with LSTM in the second model. The training of each model is based on the bagging technique to handle imbalanced data. The two models are integrated using a fusion classifier. The last step is to verify the classification result of the fusion classifier using another deep learning model to reduce the false positive. The experimental results show the superior performance of the proposed method compared to the state-of-the-art methods. In future work, we intend to improve the verification network to enhance its sensitivity and maintaining a very low false-positive rate.

ACKNOWLEDGMENTS

The work is supported by the $S\hat{e}r$ Cymru COFUND Fellowship.

REFERENCES

[1] L. Biel, O. Pettersson, L. Philipson, and P. Wide, "Ecg analysis: a new approach in human identification," *IEEE T-IM*, vol. 50, no. 3, pp. 808–812, 2001.

[2] ANSI/AAMI EC57, *Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms*. Association for the Advancement of Medical Instrumentation, 1998.

[3] C. Ye, B. V. K. Vijaya Kumar, and M. T. Coimbra, "Heartbeat classification using morphological and dynamic features of ecg signals," *IEEE T-BE*, vol. 59, no. 10, pp. 2930–2941, 2012.

[4] Z. Zhang, J. Dong, X. Luo, K.-S. Choi, and X. Wu, "Heartbeat classification using disease-specific feature selection," *Computers in Biology and Medicine*, vol. 46, pp. 79 – 89, 2014.

[5] S. Raj and K. C. Ray, "Sparse representation of ecg signals for automated recognition of cardiac arrhythmias," *Expert Systems with Applications*, vol. 105, pp. 49 – 64, 2018.

[6] V. Mondéjar-Guerra, J. Novo, J. Rouco, M. Penedo, and M. Ortega, "Heartbeat classification fusing temporal and morphological information of ecgs via ensemble of classifiers," *Biomedical Signal Processing and Control*, vol. 47, pp. 41 – 48, 2019.

[7] H. Shi, H. Wang, Y. Huang, L. Zhao, C. Qin, and C. Liu, "A hierarchical method based on weighted extreme gradient boosting in ecg heartbeat classification," *Computer Methods and Programs in Biomedicine*, vol. 171, pp. 1 – 10, 2019.

[8] Philip de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ecg morphology and heartbeat interval features," *IEEE T-BE*, vol. 51, no. 7, pp. 1196–1206, 2004.

[9] A. Sellami and H. Hwang, "A robust deep convolutional neural network with batch-weighted loss for heartbeat classification," *Expert Systems with Applications*, vol. 122, pp. 75 – 84, 2019.

[10] S. S. Xu, M. Mak, and C. Cheung, "Towards end-to-end ecg classification with raw signal extraction and deep neural networks," *J-BHI*, vol. 23, no. 4, pp. 1574–1584, 2019.

[11] S. M. Mathews, C. Kambhamettu, and K. E. Barner, "A novel application of deep learning for single-lead ecg classification," *Computers in Biology and Medicine*, vol. 99, pp. 53 – 62, 2018.

[12] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[13] S. Valverde, M. Cabezas, E. Roura, S. Gonzlez-Vill, D. Pareto, J. C. Vilanova, L. Rami-Torrent, lex Rovira, A. Oliver, and X. Llad, "Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach," *NeuroImage*, vol. 155, pp. 159 – 168, 2017.

[14] E. Essa, X. Xie, and J.-L. Jones, "Minimum s-excess graph for segmenting and tracking multiple borders with hmm," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 28–35.

[15] E. Essa and X. Xie, "Automatic segmentation of cross-sectional coronary arterial images," *Computer Vision and Image Understanding*, vol. 165, pp. 97 – 110, 2017.

[16] J. Jiang, H. Zhang, D. Pi, and C. Dai, "A novel multi-module neural network system for imbalanced heartbeats classification," *Expert Systems with Applications: X*, vol. 1, p. 100003, 2019.

[17] G. B. Moody and R. G. Mark, "The impact of the mit-bih arrhythmia database," *IEEE-EMBM*, vol. 20, no. 3, pp. 45–50, 2001.

[18] A. L. Goldberger, L. A. N. 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," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.