A Spectrogram Based Novel Approach for Arrhythmia Detection with Convolutional Neural Networks

Sneha Rao
  Michelin India Pvt. Ltd

Vishwa Mohan Singh (✉ vishwa@dozee.io)
  Turtle Shell Technologies Pvt. Ltd

Siddhivinayak Kulkarni
  MIT World Peace University

Vibhor Saran
  Turtle Shell Technologies Pvt. Ltd

A Spectrogram Based Novel Approach for Arrhythmia Detection with Convolutional Neural Networks

Sneha Rao[*,a], Vishwa Mohan Singh[*,b,†], Siddhivinayak Kulkarni [c], Vibhor Saran[b]

[a] *Michelin India Pvt. Ltd., Pune, India*

[b] *Turtle Shell Technologies Pvt. Ltd., Bengaluru, India*

[c] *School of Computer Engineering and Technology, MIT World Peace University, Pune, India*

Abstract

ECG is one of the most important medical scans which is used for diagnosis of various heart related conditions and diseases. One of the most common of these is arrhythmia, which is caused by the irregularity of the heart beats. Artificial Intelligence has had a major impact in the field of vital monitoring and autonomous medical diagnosis. Therefore, a lot of work has demonstrated its effectiveness in arrhythmia detection. In this paper, we propose a method that tries to improve upon the accuracy of such models with the help of a light weight deep learning architecture that utilized 2D Separable CNN with a group of graphical representations of the ECG signals like the STFT, CWT and MFCC. Our model has achieved an accuracy of 97.41 and an F1 score of 88.20 on a processed version of the MIT-BIH dataset and takes on an average 7.93 times less calculations compared to a simple 2D Convolution model.

Keywords: Machine Learning, Deep Learning, Electrocardiography, AI in Medicine, Time Series Analysis

---

[*] Authors have contributed equally

[†] Corresponding Author

E-mail: vishwa@dozee.io

1.   Introduction

1.1   *Background*

Arrhythmia is one of the most common heart conditions that is diagnosable with Electrocardiogram [1]. These are caused by abnormally fast, slow or irregular heartbeats. A variant of arrhythmia called atrial fibrillation was mentioned in 175,362 death certificates and was the possible cause of death of 25,845 deaths [2]. They are identified by characteristics like missing discrete P wave [3], irregularly high ventricular heart rate and the absence of iso-electric baseline [4]. Most or all of these observations can be seen in the lead II of the electrocardiogram and till some extent in lead V1 [5]. Currently, this diagnosis is done manually by doctors and technicians by looking at the ECG graph. This makes the diagnosis slow and delays the treatment the patient requires. Therefore, there is a need for methods that can assist them by automating this process.

This has led to the introduction of methods like machine learning and deep learning to help with the classification of the type of arrhythmia from the given beat. Various methods ranging from statistical machine learning to deep learning like LSTM based autoencoder [6] and Convolution Neural Network [7] has been used to classify heartbeats measured using an ECG device. However, recent advancements in the same fields have made it possible to improve the accuracy of such models with minor tweaks to the existing methodology.

1.2   *Motivation*

Most of the work in this field is done with keeping the data in a one-dimensional vector. This involves the use of methods that can encode a 1D vector into useful features while capturing the sequential trend of the signal. Although, the information given by a 1D vector is substantially small which is further affected by the low efficacy of a 1-dimensional convolution model. A better manner of representation of such signals are spectrograms, that represent the frequency variation with respect to time in a graphical format. To process such images, we need to use a 2D CNN model. However, different types of representation capture different features. Therefore, the best way to utilize the same would be to explore a method that can collectively infer from multiple such spectrograms at the same time. This is the primary motivation and goal of this work. Here, we also

explore methods and techniques that can mitigate the increased number of calculations that come with the move from a single 1D to an ensemble 2D convolution network.

The rest of the paper is organized as follows: Section 2 details the Literature Review, Section 3 describes Research Methodology, Section 4 discusses the results obtained from the experiments and Section 5 concludes the work while indicating a future direction.

2.   **Literature Review**

2.1   *Overview*

Some of the earliest works done in arrhythmia detection with the help of machine learning, like the one presented in [8], used algorithms including Random Forest, SVM and Gradient Boosting to detect arrhythmia episodes. Since then, the advancements in deep learning have helped in improving the accuracy and feature extraction methods in the same.

In [9], the authors have made a 5-layer CNN model with Exponential Linear Unit and Batch Normalization layers. The authors achieved an accuracy of 93.6% and loss of 0.2 on MIT-BIH database. The authors of [10] have made a 34-layer convolutional neural network consisting of 16 residual blocks with 2 convolutional layers per block. They were able to exceed the average cardiologist performance in both recall and precision. Authors of [11] have proposed an 11-layer CNN model consisting of 5 residual blocks with 2 convolutional layers per block. The authors evaluated the network on PhysioNet's MIT-BIH and PTB diagnostics datasets and achieved an accuracy of 93.4%.

These methods explicitly focused on using 1D models to process the signal directly. Some works try to bring this to 2D by attempting to classify the graphical representation of the same. In [12], authors of the paper have made an 11-layer 2D CNN model for arrhythmia classification and achieved an average accuracy of 99.05% on 7 heartbeat classes present in MIT-BIH arrhythmia database by converting the signals into line graph images. A better way of this is the use of spectrograms, which can highlight features much more effectively. Work done in [13] and [14] shows the effectiveness of the use of spectrograms such as Mel Spectrogram, STFT and CWT to detect and classify any arrhythmia episodes.

Based on the literature survey, it has been observed that most of the work done in arrhythmia detection and classification on ECG involves the use of 1D model. This shortcoming can be tackled with the use of 2D models. Several works have tried to do the same with a variety of inputs ranging from basic graphs to spectrograms. However, the efficacy of these models can be improved with the help of deep learning centric ensemble techniques, which have not been explored thoroughly in most of the existing research work. Along with this, most papers working with 2D CNN models fail to address the increased amount of computation that arises with the transition. Therefore, there's a need of a work that addresses both the issues, preferably with a single model.

2.2   *Objective*

The goal of this work is to develop a model that uses an ensemble of spectrograms [15] to make a model that can infer the arrhythmia episode from a variety of features to improve the prediction capabilities of the model. We also aim to overcome some of the computational overhead that comes with the use of an ensemble model with 2D CNN. This includes a method to reduce the number of features that are needed to be backpropagated along with reducing the number of calculations required to generate the output, which will help reduce both the training and the inference time of the model.

3.   **Methodology**

3.1   *Data Preprocessing*

The signal being used here in order to detect and classify arrhythmia is taken from the lead 2 of an ECG device with a sampling rate of 125 Hz. The first step in processing the signal is to remove any extraneous noises that might arise from electrical interference, breathing motion and other such sources. For this, we have used Butterworth filter [16] to only keep the frequencies within the range of 1Hz to 25Hz in the signal. The difference in the output is shown in the following Figure 1.
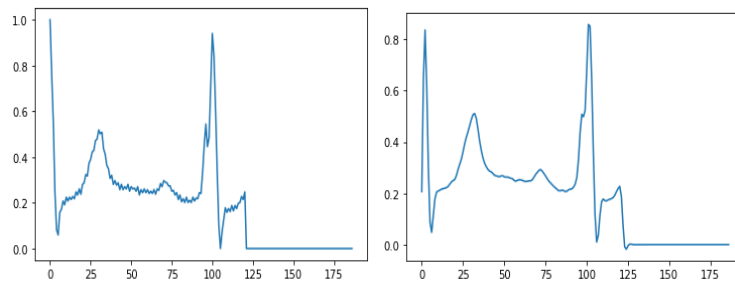
Figure 1. Raw Signal (Left) and the Filtered Signal (Right). The Butterworth bandpass filter removes the noise that lies beyond the maximum heart beat frequency of 25 Hz and the baseline created due to breathing which lies bellow 1 Hz.

The filtered signal was then converted into the graphical formats that we required for our model. For this work, we have selected 3 representations. They are as follows:

1)   *Short Term Fourier Transform*: STFT [17] uses the power vs frequency graph extracted after applying the Fourier transform [18] on small time segments of the signal to create a time vs frequency spectrogram where the color intensity opposite to a frequency reading on the y axis depends on the corresponding power in the power vs frequency curve. Here, we have taken 50 for the length of the windowed signal.

2)   *Continuous Wavelet Transform*: A CWT [19] is a convolutional transform calculated using Fourier fast transform [20] with a set of functions generated by the mother wavelet. For a mother wavelet ψ, the wavelet transform is given by:

$$W(a,b) = \int_{-\infty}^{\infty} x(t) \frac{1}{\sqrt{|b|}} \psi\left(\frac{t-a}{b}\right) dt$$

where a is the dilation parameter and b is the location parameter. We have used the Morlet wavelet [21] to create the wavelet transform.

3) *Mel Frequency Cepstral Coefficient:* MFCC [22] is a set of features that briefly represents the shape of the short-term power spectrum of a wave, based on a cosine transform of the same on a Mel scale frequency. Similar to the STFT, we have taken 50 for the windowed signal length.

All the three representations are individually resized to a shape of 128x128x3. They'll be fed to a network parallelly by concatenating all the three images into one along the 3rd axis, making the final shape of the input 128x128x9. The following Figure 2 shows the spectrogram of one of the beats extracted from the dataset.
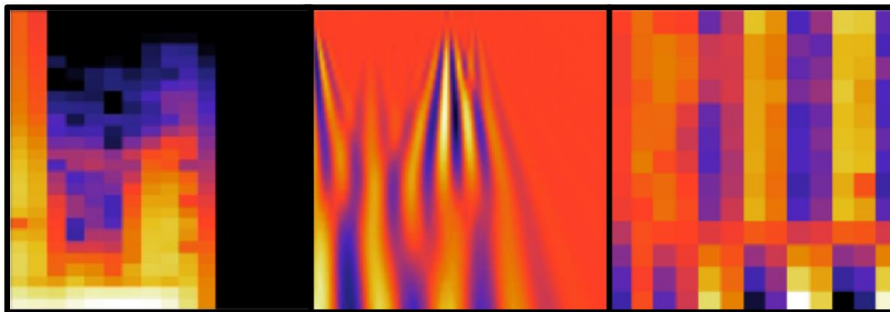
Figure 2. STFT (left), CWT (middle) and the MFCC (right) of the selected heartbeat

3.2 *Model*

The basis of the model is a general convolution 2D network. This breaks the input image into multiple lower resolution feature maps that have more focused features than the input image. Our model in total has 7 convolution layers in the backbone. However, the transition from the conventional 1D CNN to a 2D CNN with the input image having 9 layers instead of the conventional 3, increases the amount of computation required. Therefore, to first reduce the number of

calculations, we have replaced the conventional 2D CNN block with Separable 2D Convolution [23]. In a separable block, the convolution is done in 2 steps: In the depthwise operation, the block convolves each layer individually (unlike grouping them together like a conventional block) with our given kernel and feature map specifications. After this, the pointwise convolution operates on the output of the depthwise operation with a 1x1 kernel applied across all the layers in order to combine the results. This benefits the operation by reducing the number of multiplications required during such operations.

For a normal 2D convolution operation, the number of multiplications required are given by:

$$Multiplications\ (CNN) = n \times B^2 \times k^2 \times C$$

And for a subsequent separable convolution operation, the number of multiplications required are:

$$Multiplications\ (Sep\ CNN) = B^2 \times (k^2 + n) \times C$$

Where BxB is the resolution of the output image, n is the required number of feature maps, k is the size of the kernel and C is the number of channels in the input image. This makes the ratio:

$$\frac{Multiplications\ (CNN)}{Multiplications\ (Sep\ CNN)} = \frac{n \times B^2 \times k^2 \times C}{B^2 \times (k^2 + n) \times C} = \frac{n \times k^2}{n + k^2}$$

To make the convergence rate faster, we have augmented the convolution blocks with a batch normalization operation [24] followed by an ELU activation to tackle any possible dying ReLU problem. The figure 3 (a) shows the architecture of the base model.

The following Table I shows the computational advantage the separable convolution holds over the normal convolution network in the model shown in Figure 3. All the kernels have a size of 3x3.

Table I: Calculation Comparison of CNN and Separable CNN

Feature Map Count	Output Dimensions	CNN	Sep. CNN	Advantage (Sep. CNN/ CNN)
32	126x126	13716864	1952748	7.02439
32	124x124	141705216	20173312	7.02439
64	60x60	66355200	8409600	7.89041
64	58x58	124010496	15716608	7.89041
128	27x27	53747712	6391872	8.40875
128	25x25	92160000	10960000	8.40875
1024	10x10	117964800	13222400	8.92158

Secondly, to first tackle the calculations over 9 filters, we have used group convolution [25] of order 3 instead of the conventional variant. Group Convolution splits the features maps into n groups and a separate convolution block is assigned for the processing of each group. This eradicates the redundancy of the convolution operation, therefore reducing the number of connections between the layers along with the number of trainable parameters of the model. This also helps with the regularization in the case where each input group represents different features or are altogether different images by not convolving across them at the same time. In our case, we have grouped the first 6 blocks and have combined their feature maps using a 1024 feature map normal 2D Separable CNN. The same can be seen in figure 3 (b).
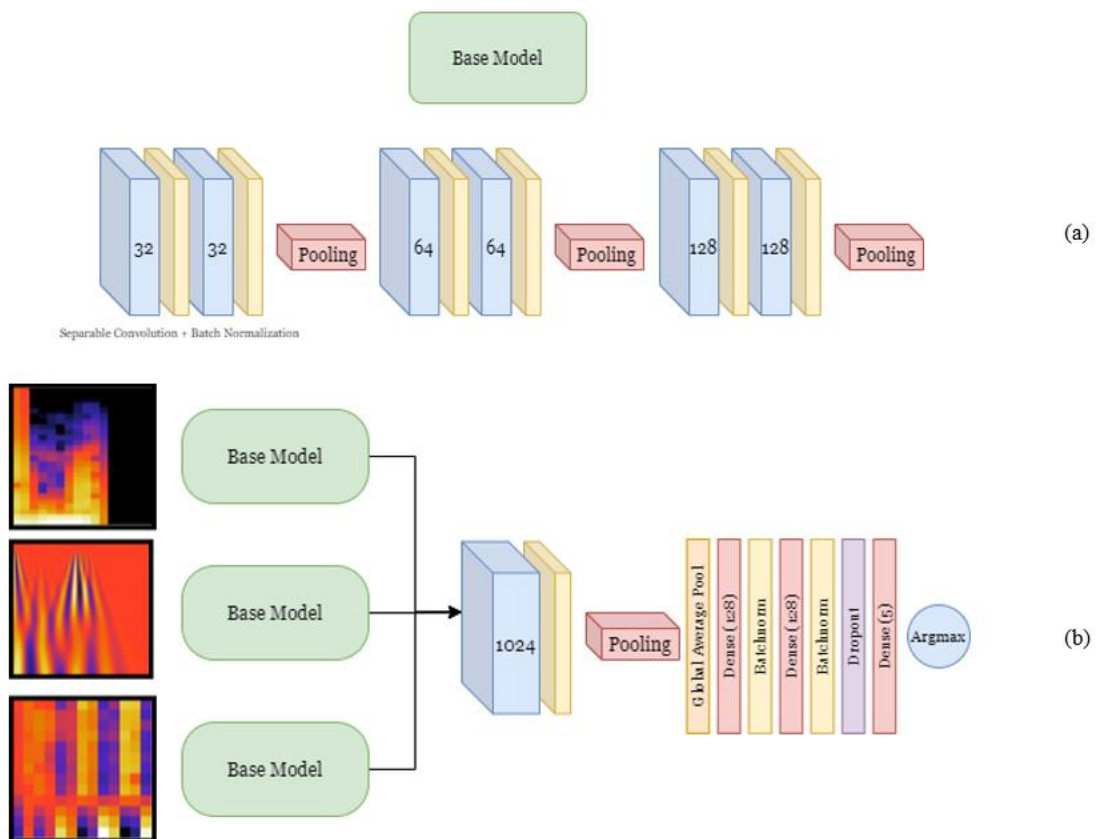
Figure 3. (a) The base model that consist of separable convolution followed by a batch normalization operation and an ELU activation. (b) The main model that uses the base model in a group convolution format and combines the output with a 1024 separable convolution block. The maps are flattened by global average pooling which is followed by a head with 2 dense layers and a dropout [27] before the 5 node SoftMax layer [28] and the argmax operation to get the output.

4.   **Result and Analysis**

4.1   *Dataset*

To evaluate our work, we have used a modified version of the MIT-BIH dataset whose preprocessing steps were described in [11]. The original MIT-BIH dataset consists of 14 classes with the signal being recorded at 360 Hz [26]. The steps in [11] groups the 14 classes into 4 broad ones and reduces the sample rate to 125 Hz. The data from lead II is taken in this case. Here, every class is associated with one heartbeat. A beat is extracted by:

- Finding the R-Peaks in the signal.
- Calculating the mean R-R distance for 10 seconds.
- Taking a sample from R-peak to R-peak + mean R-R distance with a 0 padding up to a length of 187.

To balance the classes and reduce the training time, we have reduced the number of samples belonging to class 0 in the training set to 10,000 from the original 87,554.

4.2   *Performance Comparison*

To compare our model (referred to as Spectrogram Group-Conv 2D) with the existing architectures, we have used various metrics to compare the performance. The main metric has been accuracy [29] which is the ratio of the total number of correct predictions and the total entries in the dataset. Accuracy is given by:

$$Accuracy = \frac{TP \ + \ TN}{TP \ + \ FP \ + \ TN \ + \ FN}$$

where TP and TN are true positives and negatives, and FP and FN are false positives and false negatives respectively. Precision [30] gives us an idea of the number of relevant instances from the extracted batch. Similarly, recall [30] is the fraction of relevant instance that were extracted from the dataset. They are given by:

$$Precision \ = \frac{TP}{TP \ + \ FP}$$

$$Recall \ = \frac{TP}{TP \ + \ FN}$$

To get a combined score from precision and recall, the F1 score [31] is taken that is calculated by taking the harmonic mean of the precision and recall scores. This is given by:

$$F1\ Score\ =\ \frac{2 . Precision . Recall}{Precision\ +\ Recall}$$

The following Table II compares the result of our model with existing architectures like K-nearest neighbors, 5-layer CNN [8], 34-layer CNN [9], 11-layer 1D Convolution with residuals [10] and 11-layer 2D CNN [11].

Table II: Performance comparison

Model	Accuracy	F1 score	Precision	Recall
5-layer CNN [8]	95.39	82.55	77.21	91.16
34-layer CNN [9]	96.9	85.25	80.00	92.00
11-layer 1D Convolution [10]	95.85	81.15	75.91	92.88
11-layer 2D CNN [11]	89.31	89.57	89.31	91.69
Base Model 1D	96.42	86.60	92.20	82.80
Spectrogram Group-Conv 2D	97.41	88.20	87.00	90.40
KNN [32]	95.68	82.17	81.02	83.44

The following Table III shows the class wise precision and recall for group convolution model and the Figure 4 shows the confusion matrix of the same.

Table III: Class Wise Precision and Recall

Class	Precision	Recall
Normal	0.98	0.99
Supra-ventricular	0.77	0.80
Ventricular	0.98	0.85
Fusion	0.63	0.89
Unknown	0.99	0.99

Figure 4. Confusion matrix form the group-convolution model

From the above tables and the matrix, it is evident that our model outperformed all the existing architectures in terms of accuracy when tested on an altered version of the MIT-BIH dataset. It also shows a pretty good balance between the precision and the recall score compared to some which are leaning heavily on one side.

The 2D-CNN model trained on the plain graph although outperformed every model in terms of F1-score, its low accuracy hinders the effectiveness of the model and makes it impractical to be used as a reliable source for diagnosis.

5.   **Discussion**

As the results indicate, the method in this paper obtains both its goals of improving accuracy and speed of the model. We achieved an accuracy of 97.41% on the altered version of the MIT-BIH dataset. The results achieved were better or comparable to the existing architectures that we have tested in the same environment. This can be accredited to the use of multiple spectrograms in the input.

Along with that, the number of calculations required for the classification of these spectrograms have been reduced significantly with the help of Separable convolution with group convolution also helping by reducing the number of trainable features present in the network. As compared to the conventional CNN, the separable CNN reduces the number of multiplication required in each layer by around 7-8 folds. At the same time, the group convolution makes reduces the input depth that is being fed to a CNN layer by 3 times. The presence of batch normalization and dropout along with the group convolution architecture also helps with the regularization of the model by preventing it from overfitting and boosting the convergence rate during the training process.

6.   **Conclusion and Future Scope**

In this paper, we have introduced a novel approach of analyzing ECG signals with machine learning to detect any anomalies. The efficacy of this model was shown on the MIT-BIH set for classifying the beats into 5 types of arrhythmias.

The main goal of such models is not to replace doctors or technicians, but rather to assist them in diagnosis from ECG readings so that the process of diagnosis and decision making can be

accelerated. Due to its high accuracy and low weight and inference time, this model is suitable for deployment in an environment that receives a large number of requests. This can further be improved with a deeper architecture which still manages to keep the computation required to a reasonable limit. Replacement or modifications to the input spectrograms can also help in improving the performance. This method can also be applied to other episodes that can be found in an ECG recording such as infarctions and ischemia.

7. **Declaration**

7.1 *Conflict of Interest*
The authors declare no conflict of interest.

7.2 *Acknowledgement*
We have no financial or equipment supports/funding for this project.

7.3 *Author Contributions*
S.R. and V.M.S. decided on the overall problem statement and the solution. S.R. tested the existing methods to obtain the results used in the comparison. V.M.S. identified the spectrograms and prepared the input data. S.R. and V.M.S. created the group convolution model for the input. S.R. and V.M.S. wrote the initial draft of the manuscript. S.K. provided the structure of the manuscript and outlined the content and the format. V.S. provided guidance in signal processing and classification of ECG. S.K. and V.S formatted the manuscript. All authors reviewed the manuscript.

7.4 *Ethics and consent to participate*
Not Applicable.

7.5 *Consent for publication*
Not Applicable

7.6 *Availability of data and materials*
The raw MIT-BIH dataset used in this work is open source and available at https://physionet.org/content/mitdb/1.0.0/. The processed version of the same set, which has been used in this study is also open source and available at https://www.kaggle.com/shayanfazeli/heartbeat.

References

[1] Thakor, Nitish V., and Y-S. Zhu. "Applications of adaptive filtering to ECG analysis: noise cancellation and arrhythmia detection." IEEE transactions on biomedical engineering 38.8 (1991): 785-794.

[2] Centers for Disease Control and Prevention, National Center for Health Statistics. Underlying Cause of Death, 1999–2018. Accessed March 12, 2020.

[3] Lawrence Rosenthal. "What ECG findings confirm atrial fibrillation (Afib) (AF)?". Nov 18th, 2019.

[4] Ed Burns and Robert Buttner. "Atrial Fibrillation." Jun 6, 2021.

[5] Liu SH, Cheng DC, Lin CM. Arrhythmia identification with two-lead electrocardiograms using artificial neural networks and support vector machines for a portable ECG monitor system. Sensors (Basel). 2013 Jan 9;13(1):813-28. doi: 10.3390/s130100813. PMID: 23303379; PMCID: PMC3574706.

[6] Yildirim, Ozal, et al. "A new approach for arrhythmia classification using deep coded features and LSTM networks." Computer methods and programs in biomedicine 176 (2019): 121-133.

[7] LeCun, Yann, and Yoshua Bengio. "Convolutional networks for images, speech, and time series." The handbook of brain theory and neural networks 3361.10 (1995): 1995.

[8] Batra, Anish, and Vibhu Jawa. "Classification of arrhythmia using conjunction of machine learning algorithms and ECG diagnostic criteria." Training Journal (1975).

[9] Rajkumar, A. & Marimuthu, Ganesan & Lavanya, R.. (2019). Arrhythmia classification on ECG using Deep Learning. 365-369. 10.1109/ICACCS.2019.8728362.

[10]Rajpurkar, Pranav & Hannun, Awni & Haghpanahi, Masoumeh & Bourn, Codie & Ng, Andrew. (2017). Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks.

[11]Kachuee, Mohammad & Fazeli, Shayan & Sarrafzadeh, Majid. (2018). ECG Heartbeat Classification: A Deep Transferable Representation.

[12]Jun, Tae Joon & Nguyen, Hoang & Kang, Daeyoun & Kim, Dohyeun & Kim, Young-Hak & Kim, Daeyoung. (2017). ECG arrhythmia classification using a 2-D convolutional neural network (Submitted). Submitted.

[13] Y. Byeon, S. Pan and K. Kwak, "Ensemble Deep Learning Models for ECG-based Biometrics," 2020 Cybernetics & Informatics (K&I), 2020, pp. 1-5, doi: 10.1109/KI48306.2020.9039871.

[14] J. Huang, B. Chen, B. Yao and W. He, "ECG Arrhythmia Classification Using STFT-Based Spectrogram and Convolutional Neural Network," in IEEE Access, vol. 7, pp. 92871-92880, 2019, doi: 10.1109/ACCESS.2019.2928017.

[15] Kingsbury, Brian ED, Nelson Morgan, and Steven Greenberg. "Robust speech recognition using the modulation spectrogram." Speech communication 25.1-3 (1998): 117-132.

[16] I. W. Selesnick and C. S. Burrus, "Generalized digital Butterworth filter design," in IEEE Transactions on Signal Processing, vol. 46, no. 6, pp. 1688-1694, June 1998, doi: 10.1109/78.678493.

[17] Allen, Jonathan. "Short term spectral analysis, synthesis, and modification by discrete Fourier transform." IEEE Transactions on Acoustics, Speech, and Signal Processing 25.3 (1977): 235-238.

[18] Bracewell, Ronald Newbold, and Ronald N. Bracewell. The Fourier transform and its applications. Vol. 31999. New York: McGraw-Hill, 1986.

[19] Aguiar-Conraria, Luís, and Maria Joana Soares. "The continuous wavelet transform: Moving beyond uni-and bivariate analysis." Journal of Economic Surveys 28.2 (2014): 344-375.

[20] Nussbaumer, Henri J. "The fast Fourier transform." Fast Fourier Transform and Convolution Algorithms. Springer, Berlin, Heidelberg, 1981. 80-111.

[21] Lin, Jing, and Liangsheng Qu. "Feature extraction based on Morlet wavelet and its application for mechanical fault diagnosis." Journal of sound and vibration 234.1 (2000): 135-148.

[22] On, Chin Kim, et al. "Mel-frequency cepstral coefficient analysis in speech recognition." 2006 International Conference on Computing & Informatics. IEEE, 2006.

[23] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.

[24] Santurkar, Shibani, et al. "How does batch normalization help optimization?." Proceedings of the 32nd international conference on neural information processing systems. 2018.

[25] Wang, Xijun, et al. "Fully learnable group convolution for acceleration of deep neural networks." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[26] Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209)

[27] Baldi, Pierre, and Peter J. Sadowski. "Understanding dropout." Advances in neural information processing systems 26 (2013): 2814-2822.

[28] Bridle, John S. "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters." Advances in neural information processing systems. 1990.

[29] Yin, Ming, Jennifer Wortman Vaughan, and Hanna Wallach. "Understanding the effect of accuracy on trust in machine learning models." Proceedings of the 2019 chi conference on human factors in computing systems. 2019.

[30] Davis, Jesse, and Mark Goadrich. "The relationship between Precision-Recall and ROC curves." Proceedings of the 23rd international conference on Machine learning. 2006.

[31] Fujino, Akinori, Hideki Isozaki, and Jun Suzuki. "Multi-label text categorization with model combination based on f1-score maximization." Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II. 2008.

[32] J. Park, K. Lee and K. Kang, "Arrhythmia detection from heartbeat using k-nearest neighbor classifier," 2013 IEEE International Conference on Bioinformatics and Biomedicine, 2013, pp. 15-22, doi: 10.1109/BIBM.2013.6732594.