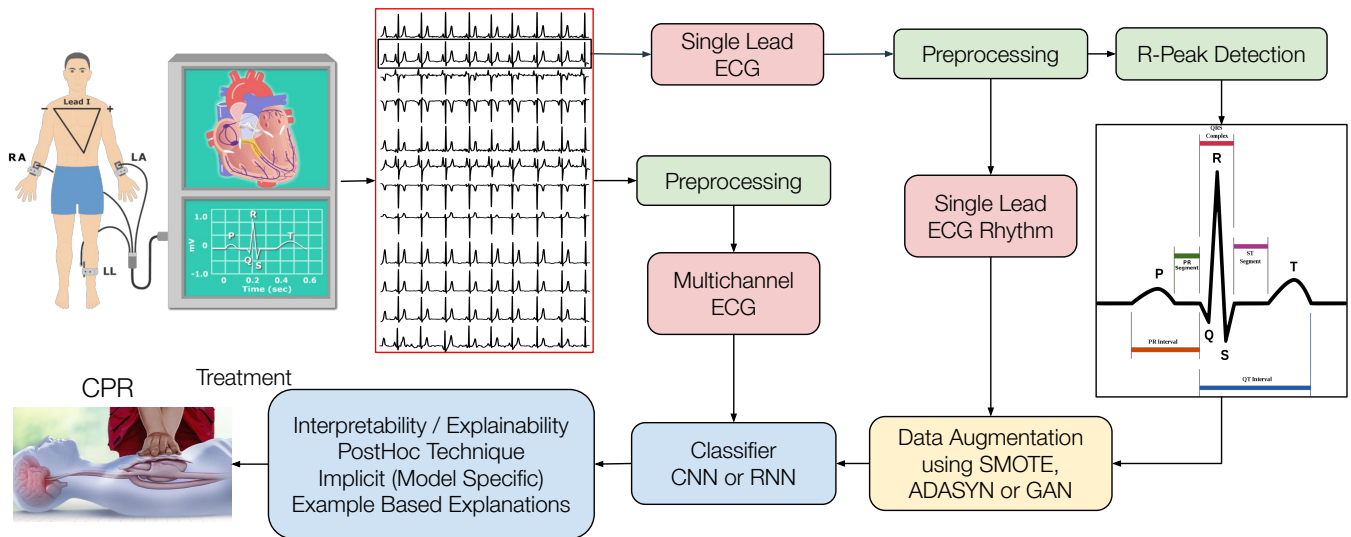


# Investigating Interpretability in Deep Learning Models: A Case Study on Cardiac Abnormality Detection from Electrocardiogram Signals

Deepankar Nankani

Department of Computer Science and Engineering  
Indian Institute of Technology Guwahati  
Guwahati, Assam, India  
d.nankani@iitg.ac.in



**Figure 1.** An Ideal ECG Classification and Interpretation Framework.

## Abstract

Availability of better healthcare services enhances life quality and prevents premature death. Cardiovascular diseases are the leading cause of natural death that can be diagnosed through an Electrocardiogram (ECG) signal. The subtle changes in ECG waveform may lead to life-threatening diseases and might be absent most of the time, making ECG classification challenging. Medical practitioners require years of training to become an expert in reading an ECG correctly. The scarcity of skilled medical practitioners necessitates

development of a Computer-Aided Disease Diagnosis System that could improve predictive healthcare and provide a less expensive yet more accurate system allowing effective patient monitoring. Cardiovascular diseases could be categorized into irregular heartbeats or rhythmic cardiac abnormalities based on their occurrence. This work investigates the aspects of interpretability in deep learning models through a case study of detecting cardiac abnormalities from single channel ECG and multi channel ECG signals. The models are further incorporated into an automated computer aided disease diagnosis system that removes noise, detects heartbeats, performs augmentation, classification, and provides explanations for the predicted diagnosis corresponding to an ECG segment as illustrated in Figure 1.

**CCS Concepts:** • Applied computing → Health care information systems; Health informatics; Bioinformatics.

**Keywords:** Deep Neural Networks, Explainable Artificial Intelligence, Healthcare, Biomedical signal processing, Electrocardiogram, Heartbeat Classification

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Doctoral Symposium, AML Systems, 2021, India

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

#### ACM Reference Format:

Deepankar Nankani. 2021. Investigating Interpretability in Deep Learning Models: A Case Study on Cardiac Abnormality Detection from Electrocardiogram Signals. In *Proceedings of Doctoral Symposium, AIML Systems '21: Doctoral Symposium, First International Conference on AI-MLSystems (AIMLSystems '21), October 21–23, 2021, Bangalore, India*. ACM, New York, NY, USA (Doctoral Symposium, AIML Systems). ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

## 1 Motivation

Life-threatening Cardiovascular diseases (CVD) are the leading cause of death annually around the world. CVD accounted for around 54.5 million deaths in 2016 in India [12]. Providing better healthcare prevents premature death and enhances the quality of life [5]. Diagnosing CVDs through electrocardiogram (ECG) signals is difficult because subtle changes in the waveform may indicate life-threatening diseases. Medical practitioners undergo years of training before performing ECG interpretation. The scarcity of expert practitioners necessitates the development of a Computer-Aided Disease Diagnosis System (CADDs). CADDs improves predictive healthcare, provides a less expensive yet more accurate and ready-to-be-deployed system for mobile devices such as smartwatches and smartphones, allowing effective patient monitoring. CVDs are associated with a pattern and can be classified as either single-channel heartbeat level cardiac abnormalities or rhythmic cardiac abnormalities that encompass a set of irregular heartbeats. We investigated both classes of cardiac abnormalities in this work.

## 2 Key challenges

The research challenges discovered during the literature survey are described subsequently.

1. **Noisy Datasets:** The low-frequency Baseline Wander, dirty or loose electrodes, might deteriorate the acquired signal quality during the data acquisition, making ECG signal classification challenging.
2. **Detecting Heartbeats:** Irregular heartbeats are prior indication of arrhythmia. Detecting the heartbeats or R-peaks is challenging as ECG is a non-stationary signal, and the signal varies from patient to patient.
3. **Synthesizing Heartbeats:** Irregular heartbeats such as supraventricular ectopic beat (SVEB) and ventricular ectopic beat (VEB) need to be detected as they lead to life-threatening arrhythmias. The classifiers require labeled, diverse, and realistic-looking heartbeats, that are difficult to acquire in practice. The challenge in synthesizing ECG occurs because the biological and physiological systems that generate these beats are highly complex. Publicly available datasets often might miss out on relevant information, thereby failing to satisfy specific criteria concerning a study.

4. **Explaining and Classifying Heartbeats:** The data generated by the population makes it difficult for cardiologists to analyze heartbeats manually. The traditional methods underperform on new datasets as their testing is limited to one or two datasets. The recent deep learning models have achieved state-of-the-art performance, but their black-box nature limits real-world deployment. In the medical domain, the explanation behind diagnosis is also necessary.
5. **Classification and Interpretation of Single Channel ECG:** CVD may appear in random episodes on the nonstationary and nonlinear ECG signals. Diseases such as Atrial and Ventricular Fibrillation occur in an episodic fashion. They are related to rapid irregular contractions of the heart's muscle fibers which are generated due to untreated SVEB and VEB beats. Therefore, classification of rhythmic arrhythmias using single channel ECG is challenging. In addition, the explanation of the diagnosis is also required to support the model decisions.
6. **Multichannel Multilabel ECG Classification and Interpretation:** Single-channel ECG provides lower resolution and therefore misses out on several cardiac abnormalities. The multi (12,6,4,3,2) channel ECG provides better resolution by capturing wide range of diagnostic information from different angles of human body. In addition, the cardiac abnormality interpreted by one cardiologist may differ when interpreted by another cardiologist, leading to multiple cardiac abnormalities for a single recording. The classification studies performed in the past are trained, tested, or developed in single, small, or relatively homogeneous datasets. The previously developed algorithms focus on identifying small numbers of cardiac arrhythmias that do not represent the complexity and difficulty of ECG interpretation.

## 3 Thesis contributions

The thesis investigates interpretability in deep learning models for detecting cardiac abnormalities. The developed models are incorporated into CAADS, which removes noise, performs augmentation and classification, and explains the diagnosis for an ECG segment. During the ECG signal acquisition, the recordings get contaminated with noise which makes ECG signal classification challenging. Therefore the low-frequency noise, baseline wander, is removed from ECG signals using Variational Mode Decomposition (VMD), a signal decomposition technique [7]. The clean ECG is used to detect irregular heartbeats or rhythmic cardiac abnormalities using deep learning models. The predicted diagnosis are later explained to stakeholders using interpretability techniques. The thesis consists of three major contributions. The first contribution is geared towards synthesis, classification, and

explanation of detected heartbeats from single channel ECG signal. The second contribution is oriented towards single channel single label ECG rhythm classification and interpretation. The third and last contribution builds upon the second contribution and classifies Multichannel Multilabel ECG rhythm classification followed by the interpretation of diagnosis. The tasks performed during each contribution are described subsequently. Section 3.1 describes the work performed for the thesis and Section 3.2 provides an idea of the work planned for the thesis.

### 3.1 Current Research

The **first** contribution aims at detecting the heartbeats from clean single channel ECG followed by augmenting, classification, and explanation of detected heartbeats. Initially, a simple, reliable, and intuitive algorithm is presented that extracts R-peaks from clean single-channel ECG using mathematical morphological operators. The operators are implemented using dynamic programming with memoization, which helps achieve accurate results in a shorter duration. The heartbeats are segmented using the extracted R-peaks from the single-channel clean ECG signal. The irregular heartbeats, namely, supraventricular ectopic beats (SVEB) and ventricular ectopic beats (VEB) are augmented using a deep convolution conditional generative adversarial network. The quality of generated heartbeats is estimated quantitatively through five evaluation metrics and qualitatively through visual representation [8].

The original and synthesised heartbeats are classified using the proposed Penalty Induced Prototype-based eXplainable Residual Neural Network (PIPxResNet) that addresses the black-box nature of deep neural networks [9]. PIPxResNet encodes the temporal variations of heartbeats by employing a pretrained residual neural network following the concept of task transfer learning. The algorithm then extracts prototypes that are most representative of the training dataset. The prototypes of a particular class having a close resemblance to other class prototypes are penalized, and their contribution towards the corresponding class is reduced. The model provides the prototypes as explanations to justify model predictions to general physicians, making the prototypes clinically relevant. The PIPxResNet model achieves significant performance, circumvents the black-box nature of the deep neural networks, and provides a prediction rationale for real-world deployment.

The cardiac abnormalities also arise rhythmically, therefore the **second** contribution is oriented towards single channel single label ECG rhythm classification and interpretation. Atrial Fibrillation (AFib) is the most commonly occurring rhythmic cardiovascular disease that causes cerebral apoplexy, stroke, and even death. The World Health Organization reported that around six million people in the United States and around ninety million people worldwide suffer

from AFib [4]. Therefore, an end-to-end framework is developed for detecting AFib from single channel ECG segments using convolution and attention based models. In addition, the interpretability is provided through attention based models by highlighting the relevant characteristic waves responsible for AFib. Part of this work is published [6] and rest is communicated.

The later part of single channel single label ECG rhythm classification and interpretation focuses on Ventricular Tachyarrhythmias such as Ventricular tachycardia (VT) and Ventricular fibrillation (VF). VT and VF accounted for 4.5% of the cardiac arrhythmia and heart failure patients in India in 2016 [16]. We detected VT-VF using a Residual Neural Network and explain the predictions using post-hoc gradient backpropagation-based techniques. The gradient backpropagation techniques investigated are Guided Backpropagation [15], Grad CAM [14], and their hybrid Guided Grad CAM [14] that highlight the signal timestamps responsible for a particular diagnosis. The highlighted timestamps might reflect peaks and convey incorrect information or explanation to the physicians. Therefore, the techniques were verified using the sanity checks [1] to verify whether they highlight relevant signal timestamps. The sanity checks describe that neural network trained weights and the input training data should affect the saliency maps. The sanity check is performed through the Neural network weight Randomization Test and Data Randomization Test [1]. The techniques passing these two tests provide correct interpretation to the stakeholders.

The **third** and last contribution classifies Multichannel Multilabel ECG rhythm classification [3, 11] followed by the interpretation of diagnosis. Initially, the Multichannel ECG is classified into a single cardiac abnormality using convolution neural networks (CNN), recurrent neural networks (RNN), combination of CNN and RNN, RNN with attention mechanism [2], and combination of CNN, RNN, and Attention Mechanism. The models underperform as only single label for each ECG segment was considered. The initial results got published in IEEE Computing in Cardiology (CinC) 2020 [10].

The Multichannel ECG (MECG) classification is further extended to incorporate demographic and heartbeat features with MECG to classify multiple labels for corresponding MECG segment [13] using a parallel convolution neural network - global average pooling (CNN-GAP) network. The demographic features include age and gender, whereas the heartbeat features include Heart Rate, RR Intervals, Mean QRS Amplitude, Hermite polynomial coefficients, statistical features, and Wave Amplitude based features. This work got accepted in IEEE CinC 2021. The extension of this work combines the aforementioned parts and provides intra-lead and inter-lead interpretability along with the classification of multi-label MECG and is currently under progress.

All the works are verified using several publicly available standard datasets obtained from multiple continents making the methods free from dataset biasing. The proposed methods can perform automated screening and provide medical attention by simulating a clinical decision support system for general physicians.

### 3.2 Future Research

In the third and last contribution we are currently adding interpretability aspects for MEEG classification through attention-based recurrent neural network, specifically, gated recurrent unit. The attention mechanism will aim at highlighting the crucial leads and signal timestamps in the corresponding leads responsible for the cardiac abnormality prediction.

## Acknowledgments

I thank my thesis advisor Dr. Rashmi Dutta Baruah.

## References

- [1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292* (2018).
- [2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
- [3] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, 23 (2000), e215–e220.
- [4] Yuki Hagiwara, Hamido Fujita, Shu Lih Oh, Jen Hong Tan, Ru San Tan, Edward J Ciaccio, and U Rajendra Acharya. 2018. Computer-aided diagnosis of atrial fibrillation based on ECG signals: A review. *Information Sciences* 467 (2018), 99–114. <https://doi.org/10.1016/j.ins.2018.07.063>
- [5] Paul Kligfield. 2002. The Centennial of the Einthoven Electrocardiogram. *Journal of Electrocardiology* 35, 4 (2002), 123–129.
- [6] Deepankar Nankani and Rashmi Dutta Baruah. 2019. An End-to-End framework for automatic detection of Atrial Fibrillation using Deep Residual Learning. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 690–695. <https://doi.org/10.1109/TENCON.2019.8929342>
- [7] Deepankar Nankani and Rashmi Dutta Baruah. 2020. Effective Removal of Baseline Wander from ECG Signals: A Comparative Study. In *International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (Communications in Computer and Information Science)*. Springer Singapore, Singapore, 310–324. [https://doi.org/10.1007/978-981-15-6318-8\\_26](https://doi.org/10.1007/978-981-15-6318-8_26)
- [8] Deepankar Nankani and Rashmi Dutta Baruah. 2020. Investigating Deep Convolution Conditional GANs for Electrocardiogram Generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8. <https://doi.org/10.1109/IJCNN48605.2020.9207613>
- [9] Deepankar Nankani and Rashmi Dutta Baruah. 2021. PIPxResNet: Penalty Induced Prototype-Based eXplainable Residual Neural Network for Heartbeat Classification. *arXiv preprint Research Square* (2021). <https://doi.org/10.21203/rs.3.rs-852812/v1>
- [10] Deepankar Nankani, Pallabi Saikia, and Rashmi Dutta Baruah. 2020. Automatic Concurrent Arrhythmia Classification using Deep Residual Neural Networks. In *2020 Computing in Cardiology (CinC)*. IEEE, 16–19.
- [11] Erick A. Perez Alday, Annie Gu, Amit Shah, Chad Robichaux, An-Kwok Ian Wong, Chengyu Liu, Feifei Liu, Bahrami Ali Rad, Andoni Elola, Salman Seyedi, Qiao Li, Ashish Sharma, Gari D. Clifford, and Matthew A. Reyna. 2020. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. *Physiological Measurement* 41 (2020). Issue 12. <https://doi.org/10.1088/1361-6579/abc960>
- [12] D Prabhakaran, P Jeemon, M Sharma, G Roth, C Johnson, S Hari Krishnan, R Gupta, J Pandian, N Naik, A Roy, et al. 2018. India State-Level Disease Burden Initiative CVD Collaborators. The changing patterns of cardiovascular diseases and their risk factors in the states of India: the Global Burden of Disease Study 1990–2016. *Lancet Glob Health* 6, 12 (2018), e1339–e1351.
- [13] Matthew A Reyna, Nadi Sadr, Erick A Perez Alday, Annie Gu, Amit Shah, Chad Robichaux, Bahrami Ali Rad, Andoni Elola, Salman Seyedi, Sardar Ansari, Qiao Li, Ashish Sharma, and Gari D Clifford. 2021. Will Two Do? Varying Dimensions in Electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. *Computing in Cardiology* 48 (2021), 1–4.
- [14] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*. 618–626.
- [15] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).
- [16] Amit Vora, Ajay Naik, Yash Lokhandwala, Arun Chopra, Jagmohan Varma, G.S Wander, Aparna Jaswal, V. Srikanthan, Balbir Singh, Dhiman Kahali, Anoop Gupta, R.R. Mantri, Anil Mishra, Ulhas Pandurangi, Debashis Ghosh, Jitendra Singh Makkar, Sujaayaa Sahu, and Rajesh Radhakrishnan. 2017. Profiling cardiac arrhythmia and heart failure patients in India: The Pan-arrhythmia and Heart Failure Observational Study. *Indian Heart Journal* 69, 2 (2017), 226–239. <https://doi.org/10.1016/j.ihj.2016.11.329>