21st Century Publisher

21 ST CENTURY CARDIOLOGY

21st Century CARDIOLOGY

Commentary

Open Access

Commentary on CEFEs, a CNN Explainable Framework for ECG Signals

Sagnik Dakshit, Prabhakaran Balakrishnan*

The University of Texas at Dallas, TX, USA

***Corresponding Author:** Professor. Prabhakaran Balakrishnan, Department of Computer Science, The University of Texas at Dallas, TX, TX 75083, USA, USA; E-mail: bprabhakaran@utdallas.edu

Received: 23 July 2023; **Accepted:** 24 July 2023; **Published:** 30 July 2023

Abstract:

In healthcare, trust, confidence, and functional understanding are critical for automated decision support systems, therefore, presenting challenges in the prevalent use of black-box deep learning (DL) models. This work attempts to summarize the global interpretable performance explanation methods developed for the task of Arrhythmia classification using deep learning. CEFEs [1] provides a task independent modular framework for investigation of learned features for 1D time-series signals in terms of shape, frequency, and clinical features. The framework allows for functional understanding of the model, quantification of model capacity, and comparison of deep learning models through interpretable explanations. The results also demonstrate understanding of model performance improvement using synthetic signals. While the modules are independent of the task, the tests for each module are task specific and demonstrated for Arrhythmia classification using ECG signals.

Keywords: *Deep learning; ECG Signals; Explainable AI; Synthetic Healthcare data*

Introduction

Recent advancements in the field of deep learning and their widespread application in daily lives has led to a strong demand for explanations to trust these black-box models. It is imperative to gain a functional understanding of these models for their deployment and acceptance in critical domains such as healthcare. While users seek to trust and gain confidence in these models [2], domain experts seek to understand the functional mechanism of black-box models with an emphasis on understanding how these models arrive at specific patient outcomes. Recent research on explainable methods such as class activation mapping [3], saliency maps [4], and GradCam [5] cannot be used as metrics for evaluation of model capacity and cannot be applied to time-series 1D data. The literature on explainable methods can be grouped into two broad categories of 1) Interpretable models [6,7] and 2) Post-hoc model explanations [3,4,5,8,9]. CEFEs [1] works towards turning black-box models into gray-box models through the latter. These explanations do not allow evaluation or quantification of learned feature space for function understanding by experts.

CEFEs:

In the CEFEs, the primary contributions of the authors can be listed as:

a) Functional understanding of the feature space of deep learning models from each of the convolution layers.

b) Understanding class discrimination learned features in terms of shape, frequency, and clinical human observable features.

c) Mapping the deep learning learned features and the features' contributions to misclassification of patient outcomes.

CEFEs is a modular framework for the quantification and visualization of the learned feature space of deep learning models in terms of shape, frequency, and clinical features. The three modules namely Descriptive Statistics, Feature Statistics, and Feature Detection and Mapping compares the learned feature maps extracted from any 1D Conv layer of the neural network model and compares it to the original ECG signal for feature mapping (Figure 1).

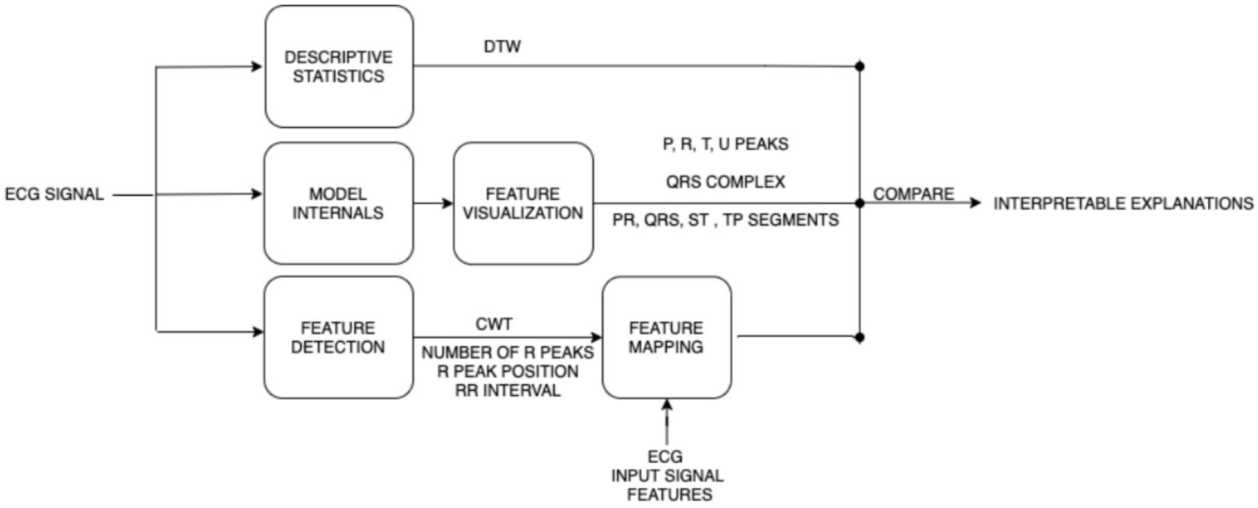

Figure 1: CEFEs modular framework.

The interpretable quantified features extracted through CEFEs' can be used to 1) Explain the quality of the deep learning model where traditional machine learning performance metrics do not suffice, 2) Explain the classification outcome of individual cases, and 3) Explain difference in performance of multiple models through feature comparison. The three modules illustrated are:

1. Descriptive Statistics: These features help realize the model's capacity to learn inherent statistical and morphological features of data samples such as shape of ECG signal. The author's use Dynamic Time Warping (DTW) algorithm to compute and quantify the similarities between the learned feature maps and the original ECG signal for each model as dtw_{intra} and can be extended to compare models as $dtw_{inter.}$

2. Features Visualization: This module promotes visual understanding of transformation of the ECG signals in the feature space through overlays of the extracted feature maps and the passed ECG signals. This visualization can become complicated and require trained medical expert intervention for fine interpretation.

3. Feature Detection and Mapping: The last module applies ECG feature detection techniques to extract clinical and frequency features from the feature maps and the corresponding input ECG signals. The authors in this paper used Continuous Wavelet Transform as the frequency feature and number of R peaks, R peak positions and the distance between corresponding R peaks in multi-beat ECG rhythm as clinical features quantified as Slack.

Discussion

The authors investigate different deep learning models trained with A% of real-world ECG data, and B% of synthetic data combined where keeping A at 100%, while the percentage of B is incremented from 0 to 100% by 20% to compare the effect of synthetic data while evaluating the CEFEs framework for understanding and comparing model quality for the task of 4-way arrhythmia classification namely Normal Sinus Rhythm, Atrial Fibrillation Peri-Ventricular Contraction, and Left Bundle Branch Block. The authors identified three cases to best evaluate model performance: a) Understanding Performance Improvement, b) Understanding Performance Degradation, c) Understanding Performance No-Change. The authors demonstrate that using a majority vote on their identified metrics of intra DTW, MSE from CWT, Slack from Clinical Features, all Arrhythmia classification outcomes can be explained from ECG signals (Table 1).

Table 1: CEFEs metric results; Intra DTW, MSE, slack represents shape, frequency and clinical features respectively. Model M<RA,SB> represents model trained with A% of real data and B% of synthetic data.

Model	Intra DTW	MSE	Slack (%)	Classification
Model Improvement				
$M<R_{100}, S_{80}>$	73.743	0.051	0.01	√
$M<R_{100}, S_{60}>$	76.248	0.118	12.11	
Model Degradation Type 1				
$M<R_{100}, S_{40}>$	874.27	0.029	4.52	√
$M<R_{100}, S_{60}>$	946.020	0.231	10.3	
Model Degradation Type 2				
$M<R_{100}, S_{0}>$	167.88	0.047	0.38	√
$M<R_{100}, S_{100}>$	317.32	0.334	3.6	
Model No Change				
$M<R_{100}, S_{0}>$	862.34	0.69	2.3	
$M<R_{100}, S_{100}>$	887.87	0.84	6.0	

Conclusion

The application of CEFEs 1) Makes black-box models trustworthy through interpretable explanations 2) Aids functional

understanding of models through interpretable explanations as feature metrics. The three modules of CEFEs namely Feature Descriptive Statistics, Feature Visualization, and Feature Detection and Mapping are independent of the task and the tests for each of the modules can be selected to fit the task criteria. The metrics can quantify and evaluate a model capacity where traditional metrics such as selectivity, sensitivity, accuracy do not sufficiently provide information about the features learned.

References

1. Maweu BM, Dakshit S, Shamsuddin R, Prabhakaran B. CEFEs: a CNN explainable framework for ECG signals. Artif Intell Med. 2021 May 1;115:102059. https://doi.org/10.1016/j.artmed.2021.102059
2. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018 Nov;19(6):1236-46. https://doi.org/10.1093/bib/bbx044
3. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. InProceedings of the IEEE conference on computer vision and pattern recognition. 2016;2921-2929.
4. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013 Dec 20. https://doi.org/10.48550/arXiv.1312.6034
5. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2019;128:336-359. https://doi.org/10.1007/s11263-019-01228-7
6. Zhang Q, Wu YN, Zhu SC. Interpretable convolutional neural networks. InProceedings of the IEEE conference on computer vision and pattern recognition. 2018; 8827-8836.
7. Sturm I, Lapuschkin S, Samek W, Müller KR. Interpretable deep neural networks for single-trial EEG classification. J Neurosci Methods. 2016 Dec 1;274:141-5. https://doi.org/10.1016/j.jneumeth.2016.10.008
8. Ribeiro MT, Singh S, Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier. InProceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016 Aug 13;1135-1144. https://doi.org/10.1145/2939672.2939778
9. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017;30.