Explainability by Shapley attribution for electrocardiogram-based algorithmic diagnosis under subtractive counterfactual reasoning setup

Arijit Ukil[1,*], Antonio J. Jara[2] and Leandro Marin[3]

[1]TCS Research, Kolkata, India
[2]Libelium, Spain
[3]University of Murcia, Spain

Abstract

Algorithmic diagnosis using Electrocardiogram (ECG) signals for various cardiovascular diseases is an important step towards developing AI-assisted healthcare systems. Explaining the predictions of algorithmic decision through machine learning models seems to be absolutely necessary for practical purposes to inculcate trust and transparency. Shapley value-based additive feature importance explanation is supported with game theoretical axioms. In this paper, we demonstrate that the Shapley value-based features do indeed directly impact the model predictability under subtractive counterfactual setup. It is validated through adversarial machine learning condition as removal-based explanations that quantify the influence of each of the inputs through simulating input removal process. We show that the model's prediction capability degradation and the model hardening with adversarial training are coupled with Shapley value attributed important features as subtractive counterfactual reasoning. Specifically, we empirically confirm that the Shapley value attributed important features put the model under lesser stress under the evasion attack and the model hardening outcome becomes more robust. We substantiate our claim with empirical results, which are demonstrated on diverse ECG data of publicly available UCR time series dataset.

Keywords

ECG, Deep learning, Algorithmic diagnosis, Explainability, Shapley value, Adversarial machine learning, Counterfactual, smart cardiac care

1. Introduction

According to WHO [1], [2] cardiovascular diseases (CVDs) are the leading cause of death globally. In 2019, it was about 32% of all deaths. 38% of the premature deaths (below the age of 70) due to noncommunicable diseases were caused by CVDs. It is also accepted that most CVDs can be prevented by addressing behavioral risk factors. Roughly 48.6% of above 20 years of age Americans have CVD [1]. Atrial fibrillation (AF) condition, a critical CVD has reached the dimension of a 21st-century epidemic with large number of reported incidents and increasing prevalence. [2]. The prevalence of CVDs, lack of treatment adherence and high-rate of later stage CVD condition detection (including life threatening Arrhythmias like Atrial Fibrillation, Ventricular Tachycardia, Atrial Flutter) highlight the urgent need for transformation of the conventional cardiac care not only due to the worldwide scarcity of cardiologists, but also to leverage the advancements in AI, sensing technologies, Internet for efficient delivery of cardiac care.

In order to develop smart cardiovascular healthcare with AI-assisted alert system as depicted in Fig 1 [3] [4], the AI-assistant needs to accurately detect the cardiovascular disease condition from the ECG

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain

*Corresponding author

arijit.ukil@tcs.com (A. Ukil); aj.jara@libelium.com (A. J. Jara); leandro@um.es (L. Marin)

[1] WHO Cardiovascular diseases factsheet https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] World Heart Report: https://world-heart-federation.org/resource/world-heart-report-2023/

sensor (mostly single-lead ECG sensor) as well as to provide explanatory basis towards the machine learning prediction such that emergency care service can be deployed which can potentially lead to reduced mortality rate and to avoid the clinical burden of delayed intervention. Further, we need to ensure trust, data security and privacy of the smart healthcare eco-system [5], [6], [7], [8]. The main motivation is to introduce AI into medical practice to speed up the clinical decision-making process of critical CVDs like Atrial Fibrillation, Myocardial Infarction, etc to enable other specialists (for e.g., primary care physicians) or the medical care givers to reliably make necessary clinical decisions using cardiac marker signal analysis (for e.g., ECG analysis) by the AI assistant for immediate clinical intervention. In fact, algorithmic diagnosis is an important component in the development of an AI assistant to treat various heart diseases. Currently, AI models or more specifically, deep learning models have shown human expert level capability of different CVD condition identification like Arrhythmias including life threatening Atrial Fibrillation condition detection using single-lead ECG signals [9], [10]. It is clinically accepted that ECG is one of the fundamental markers of cardiac health and ECG-based automated analysis and algorithmic decision pave ways towards timely diagnosis and intervention as we are experiencing severe shortage of trained cardiologists [11] [1]. In fact, convolutional neural network with skip connection-based deep learning model, proposed from Stanford claimed to provide cardiologist-level Arrhythmia condition detection capability [10].

In our context, AI-assistant is primarily a deep learning model that analyzes the single-lead ECG and generates algorithmic prediction on the plausible CVD condition of the user along with alert generation, when necessary. While accuracy of the CVD condition detection is important, the pure data-driven approach is not sufficient for the acceptance by the medical fraternity. An explanation of the results is of utmost importance [12, 13] and we need to build explainable deep learning model. There are two basic types of model explainability exists- global and local. We consider local explanation, as it is more suitable over global explanation, which considers all the statistical units among all the explanatory variables, whereas local explanation provides explanation of the explanatory variables for a focused statistical unit [14]. Shapley value-based local explanation approach (Shapley statistics was introduced in [15] and implemented in [16] as Shapley Additive explanation (SHAP) is one of the most important local explanation methods owing to its strong theoretical foundation from cooperative game theory and backed by axiomatic relevance. Shapley value-based feature attribution, a kind of additive feature attribution method, provides a single and unique solution defined under the axioms of local accuracy, consistency, and missingness [16]. Shapley value-based feature attribution is demonstrating remarkable results [17] and it is a state-of-the-art model explanation method for ECG analysis with Shapley Additive explanation (SHAP) [18, 19].

However, the model explanation's validation is not studied well to confirm the efficacy of the SHAP method. In this paper, we consider ECG analysis as the exemplary application to demonstrate that the SHAP or Shapley value-based additive feature explanation provides consistent and intuitive explanation through adversarial machine learning set up under counterfactual robustness through explanation by removal [20]. We are further motivated by CXPlain [21] that removes single or a group of inputs to measure the function's loss as the causal objective. Consequently, we follow subtractive counterfactualization. In additive counterfactuals, extra information is added to study the response, and in subtractive counterfactuals some of the given information is removed to study the response. From the understanding that additional information provisioning is an expensive exercise when medical data collection and annotation are concerned, we focus on subtractive counterfactuals. For empirical study, we experiment with different ECG data from UCR time series [22], which is the benchmark archive for time series classification problems, and we demonstrate empirical support for the consistent explanation capability of Shapley value-based explanation method.

[1] https://www.medicalindependent.ie/in-the-news/conference/ai-promises-the-gift-of-time-to-cardiologists/

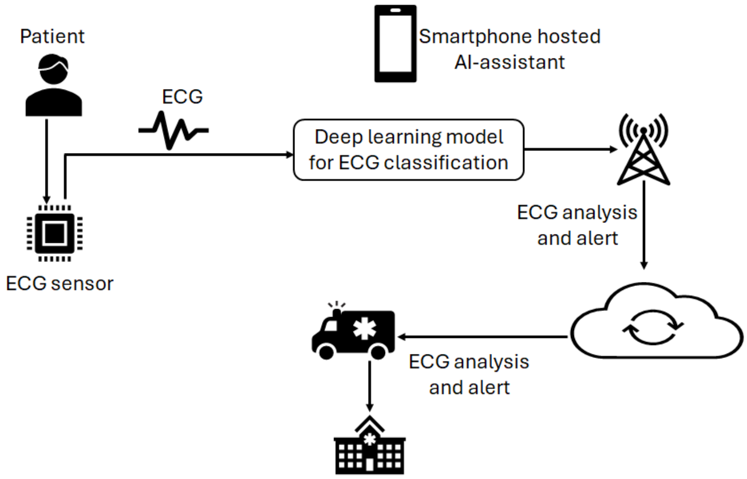
Patient
ECG
ECG sensor
Smartphone hosted AI-assistant
Deep learning model for ECG classification
ECG analysis and alert
ECG analysis and alert

Figure 1: Smart cardiovascular healthcare with AI-assisted alerting system for emergency care service delivery.

2. Problem statement, background, and solution sketch

ECG is a time series signal which is an ordered set of real values collected over time intervals and it is represented as: $\boldsymbol{x} = [x_1, x_2, x_3, ..., x_T]$, $\boldsymbol{x} \in \mathbb{R}^T$, where $\boldsymbol{x}$ consists of scalar measurements over a time period indexed by $1, 2, 3, ..., T$. Training ECG dataset $\mathbb{X}_{Train} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(N)}]$, where each of $\boldsymbol{x}^{(n)}$, $n = 1, 2, \dots, N$ consists of N number of ECG signals with corresponding labels of disease class $\mathbb{Y}_{Train}$ and the complete training dataset is D_{Train}, where $D_{Train} = [\mathbb{X}_{Train}, \mathbb{Y}_{Train}]$. Here we consider supervised learning problem to classify the given ECG signal into predicted disease class, where we construct a model $h_\theta(.)$, which is parameterized by θ describing the joint distribution $\mathcal{D}_{data}(\mathbf{x}\text{,}\mathbf{y})$ and we generate trained model M .
The first problem is to build an accurate classification model from D_{Train} under different practical challenges like less number of training examples, etc. Previously, it is demonstrated that sophisticated deep neural network model like $BlendRes^2Net$ is a suitable choice as the baseline deep neural network model [23, 17]. $BlendRes^2Net$ is a two channel blended ResNet architecture that describes the unput into both time domain and spectral domain. In general, $BlendRes^2Net$ works as a push-pull mechanism that pushes the model towards sophisticated representation with blended ResNet and pulls down to lesser network capacity with restrained learning principle. We understand the capability of a typical residual network (ResNet [24]) to minimize the vanishing gradient issue. Consequently, $BlendRes^2Net$ model provides considerable accuracy in analyzing ECG signals, but to incorporate model-level explanation, we use SHAP as the post-hoc explanation method by estimating the contribution of each of the training samples or players (player is the one who participates in the game or deal, under the game theory context) $\boldsymbol{x}^{(n)} \subset \mathbb{X}_{Train}$ towards the predictability impact of the model. To compute the Shapely value for each of the training samples, we define transferable utility game and marginal contribution from cooperative game theory concept [15, 25].

1. **(Definition I)** (Transferable utility game). A game that maps v: $2^N \rightarrow \mathbb{R}$ such that $v(\emptyset) = 0$ with the interpretation of $v(\psi)$ where ψ in 2^N, as the estimated value of coalition ψ and the value function $v(\psi)$ finds out the collective payoff for each of the player in the cooperation assumption. In our context, the model M is trained with n^{th} sample on all possible subset $\psi \subseteq 2^N$ and we estimate for each of training samples.
2. **(Definition II)** (Marginal contribution). The marginal contribution $\Delta_v(n, \psi)$ of player n with respect to the coalition ψ is defined as $\Delta_v(n, \psi) = v(\psi \cup n) - v(\psi)$.

We define Λ to be the integer permutations up to total number of given inputs (N) and $\lambda \in \Lambda$ and the predecessor set of players preceding n^{th} player in λ is represented as: $\psi_{n,\lambda} = \{m : \lambda(m) < \lambda(n)\}$. Accordingly, Shapley value $\phi_v(n)$ of n^{th} player with the function v is:

$\phi_v(n) = \frac{1}{N!}\sum_{\lambda\in\Lambda}\Delta_v(n, \psi_{n,\lambda})$.
The Shapley value $\phi_v(n)$ of n^{th} is computed through permutation logic for each of the training sample as [15]:
$\phi_v(n) = \frac{1}{N!}\sum_{\psi\subseteq\{1,2,3..,N\}}|\psi|!(N-|\psi|-1)!\Delta_v(n,\psi)$.
The training samples with higher values of $\phi_v(n)$ are the ones that contribute more to the learning of the model M [17]. Consequently, it is straightforward to assume that the model that learns without (say, top 20%) of the high contributing samples, would learn poor and that results in lesser accurate prediction. However, the response of the model M that gets trained with all the training samples including the high contributing samples (from the estimated Shapley values) and the response of the model $\mathcal{M}$ that gets trained without high contributing samples (say, with 10%, 15%, 20% removal of top contributing input samples) under adversarial machine learning conditions with evasion attack and model hardening, provide more insights on the adversarial robustness of the model M and $\mathcal{M}$. We expect adversarial robustness of M is higher than $\mathcal{M}$ and establish the counterfactual-based causal reasoning in support of the Shapley value-based model explanation, where we show that that quantum of change that requires to change the prediction is more in M than in $\mathcal{M}$. In other words, we demonstrate that the resistance towards the counterfactuals is less when the high Shapley values training samples are removed from the model training process, which directly presents the worth of Shapley value-based explanation equivalent of subtractive counterfactual based explanation with the notion of causality understanding. In ECG-based clinical diagnosis, monitoring and intervention such model-level explainability helps to build the trust for its use in practical purposes as depicted in Fig 1.

3. SHAP under conterfactual setup

Counterfactual setup is typically expressed as $p(y_x|x', y')$ which represents the probability that the outcome $Y = y$ is observed when the input is $X = x$ under the actual observation of $X = x'$ and $Y = y'$. A valid counterfactual is the one which is in fact classified as the desired class. Under a counterfactually robust classifier, the resistant to changes in X is high and the classifier attempts to classify as y, instead of y', i.e. the distance between the actual observation x and counterfactual observation x' should be high for an adversarially robust model. The Shapley value-based explanation can simulate each of the input feature not being present in the distribution $\mathcal{D}$ such that the prediction outcome is explained under the cooperative game scenario contrasting to the distribution change due to the absence of that input. Basically, Shapley value-based explanation is performed by asking a set of contrastive questions. Therefore, we encounter five types of data conditions.

- **Training dataset** D_{Train} which is the given labelled training set $D_{Train} = [\mathbb{X}_{Train}, \mathbb{Y}_{Train}]$ with $\mathbb{X}_{Train} = [\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}]$, and corresponding labels $\mathbb{Y}_{Train}$.
- **Test dataset** $\mathcal{T}_{Test}$ is the given testing dataset that is independent from the training dataset.
- **Shapley value estimated feature attributed negative dataset** D^{-Shap}_{Train} which consists of training set ($D^{-Shap}_{Train} \subset D_{Train}$) discarding the top $\alpha\%$ of the positive (important) inputs (α can be 5, 10,15,20,.., practically α is to be restricted <50).
- **Contrasting test inputs simulating counterfactuals** $\mathcal{T}^{counter}_{Test}$ which is the simulated counterfactual test inputs generated using (for e.g.) DeepFool algorithm [26].
- **Model hardening training dataset** $D^{hardening}_{Train}$ which is the adversarial training input that provides the model M (this model is trained with D_{Train}) or $\mathcal{M}$ (this model is trained with D^{-Shap}_{Train}) to train with augmentation through adversarial examples to counter the degradation of performance on $\mathcal{T}^{counter}_{Test}$.

These datasets set the stage to understand the robustness of the Shapley value-based explanation under adversarial machine learning with subtractive counterfactual set up. These datasets condition the base model M and $\mathcal{M}$. The outcomes in terms of test accuracy (acc) deliver the required understanding of the capability and robustness of the Shapley value-based explanation when measured under counterfactuals.

We consider model's classification performance loss in the absence of an input or a subset of the inputs to compute the explanations as a function and the associated outcome as: $\gamma : \mathcal{Y}(\boldsymbol{x}) \rightarrow \mathbb{R}$. The value γ that is associated with the input $\boldsymbol{x}$ indicates the behavior of the model. The following outcomes provide us the quantified idea of the Shapley value-based explanation.

- **Baseline test accuracy** is acc_{base}, which is the test accuracy when the model is trained with D_{Train} and tested over $\mathcal{T}_{Test}$ and the corresponding trained model is M.
- **Test accuracy with Shapley value estimated feature attributed negative dataset** is acc_{-Shap}, which is the test accuracy when the model is trained with D_{Train}^{-Shap} and tested over $\mathcal{T}_{Test}$ and the corresponding trained model is $\mathcal{M}$.
- **Test accuracy over counterfactuals tested with** M is $acc_{base}^{counter}$, which is the test accuracy of M when tested over $\mathcal{T}_{Test}^{counter}$.
- **Test accuracy over counterfactuals tested with** $\mathcal{M}$ is $acc_{-Shap}^{counter}$, which is the test accuracy of the model $\mathcal{M}$ when tested over $\mathcal{T}_{Test}^{counter}$.
- **Test accuracy over counterfactuals with** M **trained with model hardening training dataset** is $acc_{base-hardened}^{counter}$, which is the test accuracy of the model M hardened with adversarial training input $D_{Train}^{hardening}$ and tested over $\mathcal{T}_{Test}^{counter}$.
- **Test accuracy over counterfactuals with** $\mathcal{M}$ **trained with model hardening training dataset** is $acc_{-Shap-hardened}^{counter}$ which is the test accuracy of the model $\mathcal{M}$ hardened with adversarial training input $D_{Train}^{hardening}$ and tested over $\mathcal{T}_{Test}^{counter}$.

Our hypothesis of Shapley value-based explanation efficacy under counterfactuals is stemmed from the intuition that explanations are strongly related with the counterfactual explanations and adversarial robustness. The model M which consists of all the training inputs including the high Shapley-valued or the important ones is superior not only over the given test inputs $\mathcal{T}_{Test}$, but also over the contrasting test inputs simulating counterfactuals $\mathcal{T}_{Test}^{counter}$ than the model $\mathcal{M}$ that is trained by discarding the important inputs according to SHAP. Consequently, when the model gets hardened with augmented data $D_{Train}^{hardening}$, the response of M should be similarly better than $\mathcal{M}$ over counterfactual test inputs. More specifically, we need to establish that:

1. $acc_{base} \geqq acc_{-Shap}$,
2. $acc_{base}^{counter} \geqq acc_{-Shap}^{counter}$
3. $acc_{base-hardened}^{counter} \geqq acc_{-Shap-hardened}^{counter}$

Furthermore, consider $\mu_1 > \mu_2$, $\mu_1, \mu_2 \in \mathbb{N}^+$ and we denote $-Shap(\mu_1), -Shap(\mu_2)$ as the top $\mu_1\%$ and $\mu_2\%$ Shapley valued inputs removed in the model training. Our second hypothesis is:

1. $acc_{-Shap(\mu_2)} \geqq acc_{-Shap(\mu_1)}$,
2. $acc_{-Shap(\mu_2)}^{counter} \geqq acc_{-Shap(\mu_1)}^{counter}$
3. $acc_{-Shap-hardened(\mu_2)}^{counter} \geqq acc_{-Shap-hardened(\mu_1)}^{counter}$

While the above hypotheses are not proven, we establish our claim with empirical support over practical ECG datasets, given the model explainability is an important aspect of AI-assistive cardiac care that uses automated ECG analysis for diverse decision making and taking related actions.

4. Empirical results

In this study, we experiment with four ECG datasets, publicly available in the UCR archive, which is a benchmark dataset for timeseries classification [2] [27]. Dataset description is described in the Table 1. These datasets (Table 1) consist of separate training and testing parts. As per the general convention,

[2] https://www.cs.ucr.edu/%7Eeamonn/time_series_data_2018/

the classification performance or efficacy of the model is quantified by the test accuracy measure [28]). The training and test datasets are publicly available [27], which are standard timeseries classification benchmark data. Training-test partitioning are done by the database creator [3] .

Table 1
Experimental ECG dataset description

Dataset name	Training size	Testing size	Data length	CVD type
ECG200	100	100	96	Myocardial Infarction
ECGFiveDays	23	861	136	Change detection
ECG5000	500	4500	140	Congestive Heart Failure
TwoLeadECG	23	1139	82	Change detection

The model is developed in Python 3.5.4 on Tensorflow 1.4.0 and Keras 2.1.2 libraries. The model is trained in two Nvidia GeForce GTX 1080 GPUs of 10 GB memory with 64-bit x86 architecture, 2.60GHz clock speed 16 core Intel Xeon E5-2623 v4 CPU. For SHAP implementation, we use DeepLIFT algorithm [29] through DeepExplainer implementation [4]. To minimize the impact of non-reproducibility due to run-to-run variability owing to the non-determinism in typcal neural networks [30] [5], we experiment with more than 40 different random seeds and the reported empirical results are the highest occurring test accuracy values.

Firstly, we depict the performance of the deep neural network model for ECG classification. We design the ECG classification model following Residual Network (ResNet) architecture with novel restrained learning principle [23]. ResNet transforms the conventional layered representation learning and learns $H(\boldsymbol{x}) = F(\boldsymbol{x}) + \boldsymbol{x}$ at every layer of the network [24] so that the information in $\boldsymbol{x}$ gets a direct path to flow into the network benefiting better learnability of the model. While ResNet provides substantial advantage, we intend to ensure that the learing of ECG signals to get better as ECG is also well-defined in the spectral domain. Thus, following the $Blend - Res^2Net$ architecture [23], the deep neural network is formed with two parallel ResNet channels, where the $channel_1$ H^1 learns as:
$H^1(\boldsymbol{x}) = F(\boldsymbol{x}) + Freq(\boldsymbol{x})$
and $channel_2$ H^2 learns as:
$H^2(\boldsymbol{x}) = Freq(\boldsymbol{x}) + \boldsymbol{x}$
that ensures more detailed learning from the given ECG signals, where, $Freq(\boldsymbol{x})$ refers to the Fast Fourier Transform of input vector $\boldsymbol{x}$. These two parallel ResNet channels are merged together to constitute the final block of representation $H^{Blend}(\boldsymbol{x})$, which is followed by a Global Average Pooling layer. Cross-entropy is the loss function and softmax function is the final classification output layer. The model architecture as two parallel ResNet channels is shown in Fig 2. The model hyperparameters are inspired from [23] and we have not performed any additional hyperparameter searching methods.

Since ECG classification is part of time series classification task, we consider the baseline and state-of-the-art algorithms that are well-studied in time series classification tasks. More importantly, we choose the state-of-the-art algorithms that use UCR [27] for their experimental study forexact comparison. The state-of-the-art comparison includes 1NN-DTW-based model [31], COTE [32],time series ResNet [33], TS-Chief [34], Proximity Forest (PF) [35] and Catch22 [36]. In Table 2, the comparative study of the test accuracies from the state-of-the-art algorithms and our proposed model are shown. We can positively conclude that the proposed model performs ECG classification effectively and it is in fact a state-of-the-art model.

Next, we perform the important study to understand the empirical support to form the basis of Shapley

[3] https://timeseriesclassification.com/, https://timeseriesclassification.com/dataset.php
[4] https://github.com/slundberg/shap
[5] https://glaringlee.github.io/notes/randomness.html

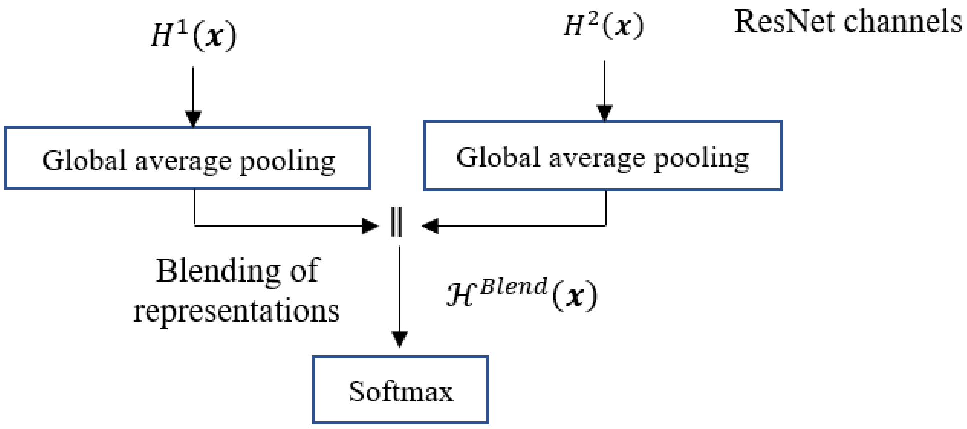

Figure 2: ECG classification model as blended ResNet architecture.

Table 2
Comparative study of test accuracies of our proposed model with relevant state-of-the-art algorithms INN-DTW([31]), COTE([32]), TS-Chief ([34]), ResNet([33]), PF([35]), Catch22([36]).

Dataset	INN-DTW([31])	COTE([32])	TS-Chief ([34])	ResNet([33])	PF([35])	Catch22([36])	Proposed
ECG200	0.88	0.88	0.855	0.8836	0.909	0.7886	0.91
ECGFiveDays	0.7967	0.9988	1.00	0.9510	84.92	0.8158	1.00
ECG5000	0.9251	0.946	0.9454	0.9510	0.9365	0.8158	0.9510
TwoLeadECG	0.86	0.993	0.9900	0.9994	0.9886	0.8539	0.9994

value-based explainability through the lens of counterfactual set up with adversarial machine learning. From the hypothesis as stated in 3, we empirically study the impact of the Shapley explanation in a counter-intuitive approach and quantitatively evaluate the hypothesis through different test accuracies $acc_{base}, acc_{-Shap(10,15,20)}, acc^{counter}_{base}, acc^{counter}_{-Shap(10,15,20)}, acc^{counter}_{base-hardened}, acc^{counter}_{-Shap(10,15,20)-hardened}$, where, $\mu_1 = 5, \mu_2 = 10, \mu_2 = 15$. The baseline test accuracy is acc_{base}, which is from the trained model M that is trained with D_{Train}. We find the test accuracies from the models with Shapley value estimated feature attributed negative datasets at different levels of removal of the input training samples (removing top 10%, 15%, 20% of the input samples respectively) denoted as $acc_{-Shap(10,15,20)}$ for the trained model $\mathcal{M}^{10,15,20}$ when the model is trained with $D^{-Shap(10,15,20)}_{Train}$. The empirical results are depicted in 3, 4, and 5. The most interesting part is the consistent degradation of the performance with more Shapley-negative training (higher amount of top input removal in the training set, i.e. with higher μ values) and the consistency is similarly observed in the counterfactual simulating contrasting testing as well as in case of model hardening. While performance degradation in Shapley-negative training indicates the impact of the Shapley value explained inputs towards the predictability of the model, the trend of lesser recovery of higher Shapley negative training (for e.g., 15%, 20%) demonstrates the support of explainability through Shapley attribution under subtractive counterfactual set up ($acc_{-Shap(10)^{counter}} \geq acc_{-Shap(15)^{counter}} \geq acc_{-Shap(20)^{counter}}$ for each of the experimental datasets as well as $acc^{counter}_{-Shap(10)-hardened} \geq acc^{counter}_{-Shap(15)-hardened} \geq acc^{counter}_{-Shap(20)-hardened}$ for each of the experimental datasets) confirming our hypothesis. More precisely, we provide empirical evidence to support the explanation process of Shapley value-based input attribution which is strongly related to counterfactuals provisioning.

We intend to mention that the application space is restricted to ECG classification due to its immediate importance in the development and deployment of smart cardiovascular system. However, we can extend the application for other relevant healthcare data analysis and classification tasks given that the model-level explainability is an utmost importance property for the acceptance of AI-assistive solutions in healthcare domains including cardiovascular care for practical purpose.

Table 3
Shapley explanation with counterfactuals with contrasting test inputs at top 10% discard of training samples for Shapley-negative training.

Dataset	acc_{base}	$acc_{-Shap(10)}$	$acc^{counter}_{base}$	$acc_{-Shap(10)^{counter}}$	$acc^{counter}_{base-hardened}$	$acc^{counter}_{-Shap(10)-hardened}$
ECG200	0.91	0.892	0.781	0.751	0.887	0.85
ECGFiveDays	1.00	0.988	0.844	0.835	0.950	0.935
ECG5000	0.9510	0.934	0.892	0.868	0.926	0.907
TwoLeadECG	0.9994	0.977	0.902	0.844	0.960	0.912

Table 4
Shapley explanation with counterfactuals with contrasting test inputs at top 15% discard of training samples for Shapley-negative training.

Dataset	acc_{base}	$acc_{-Shap(15)}$	$acc^{counter}_{base}$	$acc_{-Shap(15)^{counter}}$	$acc^{counter}_{base-hardened}$	$acc^{counter}_{-Shap(15)-hardened}$
ECG200	0.91	0.845	0.781	0.705	0.887	0.788
ECGFiveDays	1.00	0.935	0.844	0.774	0.958	0.843
ECG5000	0.9510	0.891	0.892	0.796	0.926	0.862
TwoLeadECG	0.9994	0.898	0.902	0.776	0.961	0.863

Table 5
Shapley explanation with counterfactuals with contrasting test inputs at top 20% discard of training samples for Shapley-negative training.

Dataset	acc_{base}	$acc_{-Shap(20)}$	$acc^{counter}_{base}$	$acc_{-Shap(20)^{counter}}$	$acc^{counter}_{base-hardened}$	$acc^{counter}_{-Shap(20)-hardened}$
ECG200	0.91	0.785	0.781	0.657	0.887	0.698
ECGFiveDays	1.00	0.824	0.844	0.708	0.958	0.811
ECG5000	0.9510	0.802	0.892	0.720	0.926	0.798
TwoLeadECG	0.9994	0.803	0.902	0.719	0.961	0.767

5. Conclusion

Automation in the medical care, particularly in critical care that includes cardiovascular care not only increases the efficiency in the overall medical process, but also can lead to timely intervention through AI assistant that can potentially result in lesser mortality rate and reduced clinical burden. In this paper, we anchor upon smart cardiovascular system using ECG sensor for in-home, remote and emergency care that can enable emergency cardiovascular care without delaying the life-saving intervention process. However, to embrace the automation or algorithmic clinical condition detection and alert generation for initiating required cardiovascular care, the machine learning algorithm requires to provide model-level explanation to justify the algorithmic prediction of the disease condition. Shapley value-based explanations backed by strong theoretical foundation from coalition game axioms is the apt choice and we formulate our hypothesis on the applicability of Shapley based explanations under subtractive counterfactual reasoning set up and demonstrated through empirical study on number of ECG datasets. However, we like to mention that the Shapley value computation is computationally challenging and DeepExplain with DeepLift algorithm or its variants can provide better computational efficiency. The method that we have proposed in this paper to support Shapley value-based explanations through subtractive counterfactual reasoning is generic in nature and we have chosen ResNet architecture due to its strong performance in different related classification tasks. A study with other architectures will certainly make the claim stronger. Further, in our future work, we intend to work on the theoretical basis of the proposed hypotheses and to provide more empirical evidences on related medical domains. We also intend to point out that the corresponds of the clinical explanation with Shapley value-based statistical explanation are not theoretically, hypothetically or empirically established. In this paper, our main motivation is to understand the influence of input training instances towards the model's predictability, which in turn provides quantified explanation. Future research scope certainly includes the quantified explainability with qualitative explainability with respect to relevant and standard clinical domain knowledge.

Acknowledgments

We thank and acknowledge the support and continuous encouragement of Leandro Marin, University of Murcia, Spain for this research work..

References

[1] S. S. Martin, A. W. Aday, Z. I. Almarzooq, C. A. Anderson, P. Arora, C. L. Avery, C. M. Baker-Smith, B. Barone Gibbs, A. Z. Beaton, A. K. Boehme, et al., 2024 heart disease and stroke statistics: A report of us and global data from the american heart association, Circulation 149 (2024) e347–e913.

[2] J. Kornej, C. S. Börschel, E. J. Benjamin, R. B. Schnabel, Epidemiology of atrial fibrillation in the 21st century: novel methods and new insights, Circulation research 127 (2020) 4–20.

[3] C. Puri, A. Ukil, S. Bandyopadhyay, R. Singh, A. Pal, A. Mukherjee, D. Mukherjee, Classification of normal and abnormal heart sound recordings through robust feature selection, in: 2016 Computing in Cardiology Conference (CinC), 2016, pp. 1125–1128.

[4] A. Ukil, A. J. Jara, L. Marin, Data-driven automated cardiac health management with robust edge analytics and de-risking, Sensors 19 (2019) 2733.

[5] A. Ukil, Secure trust management in distributed computing systems, in: 2011 Sixth IEEE International Symposium on Electronic Design, Test and Application, 2011, pp. 116–121. doi:10.1109/DELTA.2011.29.

[6] A. Ukil, J. Sen, Secure multiparty privacy preserving data aggregation by modular arithmetic, in: 2010 First International Conference On Parallel, Distributed and Grid Computing (PDGC 2010), IEEE, 2010, pp. 344–349.

[7] A. Ukil, L. Marin, A. J. Jara, Priv-aug-shap-ecgresnet: Privacy preserving shapley-value attributed augmented resnet for practical single-lead electrocardiogram classification, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. doi:10.1109/ICASSP49357.2023.10096437.

[8] A. Ukil, Secure trust management in distributed computing systems, in: 2011 Sixth IEEE International Symposium on Electronic Design, Test and Application, IEEE, 2011, pp. 116–121.

[9] A. Ukil, L. Marin, S. C. Mukhopadhyay, A. J. Jara, Afsense-ecg: Atrial fibrillation condition sensing from single lead electrocardiogram (ecg) signals, IEEE Sensors Journal 22 (2022) 12269–12277.

[10] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, A. Y. Ng, Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, Nature medicine 25 (2019) 65–69.

[11] W. B. Fye, Cardiology workforce: there's already a shortage, and it's getting worse!, 2002.

[12] P. Wagner, T. Mehari, W. Haverkamp, N. Strodthoff, Explaining deep learning for ecg analysis: Building blocks for auditing and knowledge discovery, Computers in Biology and Medicine (2024) 108525.

[13] Y. M. Ayano, F. Schwenker, B. D. Dufera, T. G. Debelee, Interpretable machine learning techniques in ecg-based heart disease classification: A systematic review, Diagnostics 13 (2023). URL: https://www.mdpi.com/2075-4418/13/1/111. doi:10.3390/diagnostics13010111.

[14] P. Giudici, E. Raffinetti, Shapley-lorenz explainable artificial intelligence, Expert systems with applications 167 (2021) 114104.

[15] L. S. Shapley, et al., A value for n-person games (1953).

[16] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[17] A. Ukil, L. Marin, A. J. Jara, When less is more powerful: Shapley value attributed ablation with augmented learning for practical time series sensor data classification, Plos one 17 (2022) e0277975.

[18] S. Katsushika, S. Kodera, S. Sawano, H. Shinohara, N. Setoguchi, K. Tanabe, Y. Higashikuni, N. Takeda, K. Fujiu, M. Daimon, et al., An explainable artificial intelligence-enabled electrocardiogram analysis model for the classification of reduced left ventricular function, European Heart Journal-Digital Health 4 (2023) 254–264.

[19] W. Sun, S. V. Kalmady, N. Sepehrvand, A. Salimi, Y. Nademi, K. Bainey, J. A. Ezekowitz, R. Greiner, A. Hindle, F. A. McAlister, et al., Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms, NPJ Digital Medicine 6 (2023) 21.

[20] I. Covert, S. Lundberg, S.-I. Lee, Explaining by removing: A unified framework for model explanation, Journal of Machine Learning Research 22 (2021) 1–90.

[21] P. Schwab, W. Karlen, Cxplain: Causal explanations for model interpretation under uncertainty, Advances in neural information processing systems 32 (2019).

[22] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, E. Keogh, The ucr time series archive, IEEE/CAA Journal of Automatica Sinica 6 (2019) 1293–1305.

[23] A. Ukil, A. J. Jara, L. Marin, Blend-res 2 net: Blended representation space by transformation of residual mapping with restrained learning for time series classification, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 3555–3559.

[24] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[25] A. E. Roth, The Shapley value: essays in honor of Lloyd S. Shapley, Cambridge University Press, 1988.

[26] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.

[27] H. A. Dau, E. Keogh, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, Yanping, B. Hu, N. Begum, A. Bagnall, A. Mueen, G. Batista, Hexagon-ML, The ucr time series classification archive, 2018.

[28] A. Bagnall, J. Lines, A. Bostrom, J. Large, E. Keogh, The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances, Data mining and knowledge discovery 31 (2017) 606–660.

[29] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, in: International conference on machine learning, PMLR, 2017, pp. 3145–3153.

[30] D. Zhuang, X. Zhang, S. Song, S. Hooker, Randomness in neural network training: Characterizing the impact of tooling, Proceedings of Machine Learning and Systems 4 (2022) 316–336.

[31] J. Lines, A. Bagnall, Time series classification with ensembles of elastic distance measures, Data Mining and Knowledge Discovery 29 (2015) 565–592.

[32] A. Bagnall, J. Lines, J. Hills, A. Bostrom, Time-series classification with cote: the collective of transformation-based ensembles, IEEE Transactions on Knowledge and Data Engineering 27 (2015) 2522–2535.

[33] Z. Wang, W. Yan, T. Oates, Time series classification from scratch with deep neural networks: A strong baseline, in: 2017 International joint conference on neural networks (IJCNN), IEEE, 2017, pp. 1578–1585.

[34] A. Shifaz, C. Pelletier, F. Petitjean, G. I. Webb, Ts-chief: a scalable and accurate forest algorithm for time series classification, Data Mining and Knowledge Discovery 34 (2020) 742–775.

[35] B. Lucas, A. Shifaz, C. Pelletier, L. O'Neill, N. Zaidi, B. Goethals, F. Petitjean, G. I. Webb, Proximity forest: an effective and scalable distance-based classifier for time series, Data Mining and Knowledge Discovery 33 (2019) 607–635.

[36] C. H. Lubba, S. S. Sethi, P. Knaute, S. R. Schultz, B. D. Fulcher, N. S. Jones, catch22: Canonical time-series characteristics, Data Mining and Knowledge Discovery 33 (2019) 1821–1852.