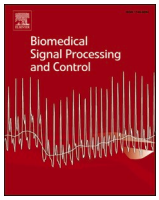
Biomedical Signal Processing and Control 75 (2022) 103584

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc



Explainable AI decision model for ECG data of cardiac disorders

Atul Anand [a,1], Tushar Kadian [a], Manu Kumar Shetty [b], Anubha Gupta [a,*]

[a] SBILab, Department of ECE, IIIT-Delhi, New Delhi, India
[b] Department of Pharmacology, Maulana Azad Medical College, New Delhi, India

ARTICLE INFO

Keywords:
Electrocardiogram
ECG Waves
Deep Learning
CNN
Residual Networks
SHAP
Interpretability
XAI

ABSTRACT

Electrocardiogram (ECG) data is used to monitor the electrical activity of the heart. It is known that ECG data could help in detecting cardiac (heart) abnormalities. AI-enabled automated analysis of ECG waves has many applications in the medical domain, such as diagnostic of heart diseases, prediction of stress level, etc. In this study, we implemented a number of deep neural networks on a publicly available dataset of PTB-XL of ECG signals for the detection of cardiac disorders. Our proposed ST-CNN-GAP-5 model produced better results compared to the existing state-of-the-art results on this dataset, achieving an AUC of 93.41%. The same network architecture is tested on another ECG dataset of arrhythmia patients to assess the generalizability of our DL model for ECG datasets, yielding an accuracy of 95.8% and an AUC of 99.46%, which is competitive in performance to the state-of-the-art models. Finally, we analyzed the ECG data using SHapley Additive exPlanations (SHAP) on the trained ST-CNN-GAP-5 to assess the explainability or interpretability of the decisions of this deep convolution network model. Results indicate that the model is able to highlight relevant alterations of the ECG waves as required by clinicians, making it explainable for diagnostic purposes. Deployment of such models can help in easing the burden on medical infrastructure in low- and middle-income populous countries.

1. Introduction

An electrocardiogram (ECG) is a graphic representation of electric potentials generated by the heart. ECG is a quick, safe, non-invasive, inexpensive and painless test commonly used for the identification of arrhythmia, conduction abnormalities, ventricular hypertrophy, and myocardial infarction [31]. Each ECG wave is labelled alphabetically, namely as P, Q, R, S, T, U (Fig. 1). P wave represents atrial depolarization. The QRS complex represents ventricular depolarization, and the ST-T-U complex (ST segment, T, and U) represents ventricular repolarization [21]. Generally, 10 electrodes are used to generate 12 conventional ECG leads, broadly divided into two groups: six limb leads (I, II, III, aVL, aVR and aVF) and six chest leads (V1, V2, V3, V4, V5 and V6). Changes in the electric potential on the frontal plane is recorded in the limb leads, and the potential in the horizontal plane are recorded in the chest leads [10].

These days, heart conduction signals can be captured via sensors tied on wrist bands while a subject is doing routine tasks. This allows automated assessment of heart functionality, and recommend strategies accordingly to help people lead healthy lives. This is further aided by energy efficient data capture techniques such as compressive sensing to support telemedicine [4]. Similarly, efficient machine learning (ML) or deep learning (DL) models can provide assistance in cardiac disease diagnosis. Since the healthcare provider would like to trust the decision of an AI model, it is helpful if these models also provide visual interpretability of the results or decision taken. Thus, today the emphasis in the area of heart conduction signal analysis is on the human centric design of wearables, accuracy of the designed AI models, and the interpretability of the AI models on the results obtained. This work focuses on the last two aspects of developing an efficient and interpretable AI model to support clinicians in the diagnosis of heart diseases.

ECG plays a critical role in early assessment in the emergency wards for triage of patients reporting with sudden chest pain [9]. ECG tests are cost-effective, affordable, and easily available in clinics, where ECG test results are also immediately available. Thus, this modality is of high clinical importance for evaluation of cardiac disorders. Although ECG machines are easily deployable in rural/urban areas, diagnosis of cardiac disorder using ECG requires fully trained cardiologists, who may not be always available in rural areas. Moreover, at the time of emergency, particularly, during odd working hours when only resident

* Corresponding author.
E-mail address: anubha@iiitd.ac.in (A. Gupta).
[1] Equal contribution.

https://doi.org/10.1016/j.bspc.2022.103584
Received 4 October 2021; Received in revised form 22 January 2022; Accepted 12 February 2022
Available online 23 February 2022
1746-8094/

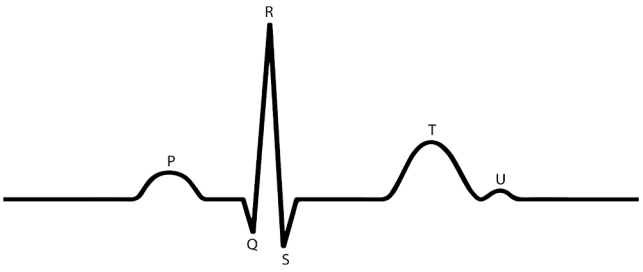


Fig. 1. PQRSTU ECG Waveform.

doctors are on duty, the diagnosis may not be accurate all the time because the resident doctors may not have the required expertise, while the time required to analyze the results is less [8]. This becomes even more challenging when the ECG data contains noise [8]. Therefore, there is an emphasis on the development of machine learning (ML) models for the automatic detection of heart diseases [17]. Diagnosis via an efficient ML model on ECG signals can save considerable time of cardiologists, while decreasing the number of misdiagnoses at the same time. Often, AI models appear as black boxes to the doctors because they do not explain the reasons as to why a particular decision is being made by these trained models [11]. Owing to this lack of interpretability/explainability of results with AI models, it is hard to deploy them in a clinical setup for decision making. Since physicians are held accountable for medical decisions [19], explainability of AI-models is essential for enabling development and deployment of cost-effective AI-based solutions in healthcare. Hence, for the diagnosis of cardiac disorders, it is required to build an efficient AI model that utilizes ECG data and presents visual interpretability of results that helps the cardiologists trust the decisions of the AI model.

1.1. Related Works

In general, the analysis of ECG data involves preprocessing of data including noise removal [1,2,33], followed by feature extraction in time domain or frequency domain using traditional signal processing techniques. These features are further utilized by various machine learning (ML) and deep learning (DL) methods for the identification of different types of cardiovascular diseases. A variety of popular and relevant techniques are discussed in subsequent subsections.

1.1.1. Feature Extraction

The traditional ML approaches involve manual extraction of features from ECG signals using the conventional feature extraction techniques or through domain knowledge [44]. Marinho et al. [26] applied feature extraction techniques such as Fourier transform (FT), Goertzel, Higher-Order Statistics (HOS), and Structural Co-Occurrence Matrix (SCM). The extracted features are fed into discriminative models to classify ECG signals. However, these models are generally not robust to noise [29]. Moreover, with manually crafted features, the overall performance depends considerably upon the quality and accuracy of features extracted because the manually selected features may not always be relevant. To address this concern, Chen et al. [12] used feature selection to select the most relevant features, where feature vectors were prepared by using wavelet transform coefficients and discrete Fourier transform spectrum extracted from the ECG signals.

Swain et al. [37] used ECG signals to identify the presence of myocardial infarction (MI) by using modified Stockwell transform (MST) based time–frequency analysis and phase distribution pattern. Alterations in ECG waveform corresponding to MI are reflected in phase distribution pattern, which is used as the discriminative features. Some authors proposed a system named SpEC to improve the classification of ECG data with a limited training dataset [3]. It involves the use of Stockwell transform (ST) to convert ECG signal into time–frequency domain and 2D-ResNet to classify the resultant images from ST into five different classes of arrhythmia.

Table 1
Some Important DL works used for ECG analysis.

Author	Technical Approaches	Cardiac Abnormalities Classified	Other Details
[37]	Modified Stockwell transform (MST) based time–frequency analysis and phase distribution pattern analysis	Identification of the presence of myocardial infarction (MI) using ECG signals	Alterations in ECG waveform corresponding to MI are reflected in phase distribution pattern, which is used as the discriminative feature.
[3]	Stockwell transform (ST) is used to convert ECG signal into time–frequency domain and 2D-ResNet is used to classify resultant images from ST	Five different classes of arrhythmia	The proposed (SpEC) system was used to improve the classification of ECG data with a limited training dataset.
[15]	Convolutional neural network (CNN) architectures such as AlexNet, VGG-16, and ResNet-18 are used.	Classification of ECG signal into two classes, namely normal and abnormal	Performed pre-processing for noise removal and feature extraction to determine the main characteristics of data. A spectrogram* was made and fed as images to CNN model as input. Results indicated that AlexNet performed the best, followed by ResNet.
[12]	Deep learning classification model (ResNet-34 with 1D convolutional layers) is used to extract discriminative features and bidirectional LSTM with one layer is used to capture temporal information of the data.	Detection of six common types of cardiac arrhythmia, including atrial fibrillation (AF), atrial flutter, complete AV block, junctional rhythm, sinus node disease and Wolff Parkinson-White syndrome (ventricular pre-excitation)	–
[7]	CNN model composed of ten residual blocks is built.	Detection of the presence of atrial fibrillation (AF) during normal sinus rhythm	CNN model worked on raw ECG data instead of signal images.
[6]	CNN model consisting of spatial and temporal layers is used for feature extraction from ECG data.	Estimation of the age and gender of a patient	After analyzing the results, they highlighted that the discrepancy between ECG age and chronological age is a marker of physiological health.
[29]	Developed deep learning models by applying CNN in conjunction with other algorithms such as KNN, Support Vector Machine (SVM), multilayer perceptron (MLP)	Identification of patients with paroxysmal atrial fibrillation (PAF)	–

*spectrogram: 2-D representation of the time–frequency information of a signal.

1.1.2. ML/DL Approaches for Classifying Cardiac Abnormalities

Popular algorithms such as fuzzy c-means [46], k-Nearest Neighbour

Table 2
List of Abbreviations.

Abbreviation	Meaning
AI	Artificial Intelligence
AUC	Area Under the Curve
CD	Conduction Disturbance
CNN	Convolutional Neural Network
DCT	Discrete cosine transform
DL	Deep Learning
DNN	Deep Neural Network
ECG	Electrocardiogram
FT	Fourier transform
GAP	Global Average Pooling
HOS	Higher-Order Statistics
HYP	Hypertrophy
kNN	k-nearest neighbour
LAFB	Left Anterior Fascicular Block
LSTM	Long Short-Term Memory
LVH	Left Ventricular Hypertrophy
MI	Myocardial Infarction
ML	Machine Learning
MLP	Multilayer Perceptron
MST	Modified Stockwell Transform
NORM	Normal ECG
PAF	Paroxysmal Atrial Fibrillation
RBBB	Right Bundle Branch Block
ResNet	Residual Networks
SCM	Structural Co-Occurrence Matrix
SENet	Squeeze-and-Excitation Network
SHAP	Shapley Additive Explanations
ST^1	Stockwell Transform
ST^2	Spatio-Temporal
STTC	ST/T Change
SVM	Support Vector Machine
WELM	Weighted Extreme Learning Machine

(kNN) [5,32], Naïve-Bayes classifier [26], Weighted ELM (WELM) [39] and Support Vector Machines (SVM) can be used for ECG signal classification. Recently, DL models have gained importance in the area of ECG signal analysis. They are generally trained on the raw data. An often used methodology in ECG research is to convert 1-D data to images. For example, Diker et al. [15] used convolutional neural network (CNN) architectures such as AlexNet, VGG-16, and ResNet-18 to classify ECG signal into two classes, namely normal and abnormal. They first performed pre-processing for noise removal and feature extraction to determine the main characteristics of data. A spectrogram (a 2-D representation of the time–frequency information of a signal) was made and fed as images to a CNN model as input. Their results indicated that AlexNet performed the best, followed by ResNet.

Attia et al. [7] developed a CNN model composed of ten residual blocks that worked on raw ECG data to detect the presence of atrial fibrillation (AF) during normal sinus rhythm. In another work, they proposed a CNN model that consisted of spatial and temporal layers for feature extraction from ECG data to estimate the age and gender of a patient [6]. After analyzing the results, they highlighted that the discrepancy between ECG age and chronological age is a marker of physiological health. Pourbabaee et al. [29] developed deep learning models by applying CNN in conjunction with other algorithms such as KNN, Support Vector Machine (SVM), multilayer perceptron (MLP) to identify patients with paroxysmal atrial fibrillation (PAF).

The work of Strodthoff et al. [35] is considered state-of-the-art because they benchmarked the performance of different models on the PTB-XL dataset and also included results of other authors. They presented benchmarking results covering a variety of models including resnet1d-wang, XResNet1d101, LSTM, fcn-wang, Inception1d, lstm-bidir, Wavelet + NN. They interpreted that CNNs, especially, ResNet and Inception-based architectures perform better than other models. The best performing model is named "resnet1d wang" that was proposed by Wang et al. [43] and adapted by Strodthoff et al. [35] for one-dimensional inputs. It stacks three residual blocks, followed by a global average pooling layer and an output layer. Some relevant ML and DL techniques are presented in Table 1.

Motivated with the above discussion, the aim of this work is to build a generalized deep learning model for supporting the diagnosis of the cardiac abnormalities. In addition, we would like to provide visual explanations or interpretability that leads to decision making by the AI model. It is desired that these interpretations are understandable by the clinicians, so that they can trust the AI model for use in actual clinical practice.

This paper is organized as follows. Section 2.1 describes two large publicly available ECG datasets that have been utilized for training and

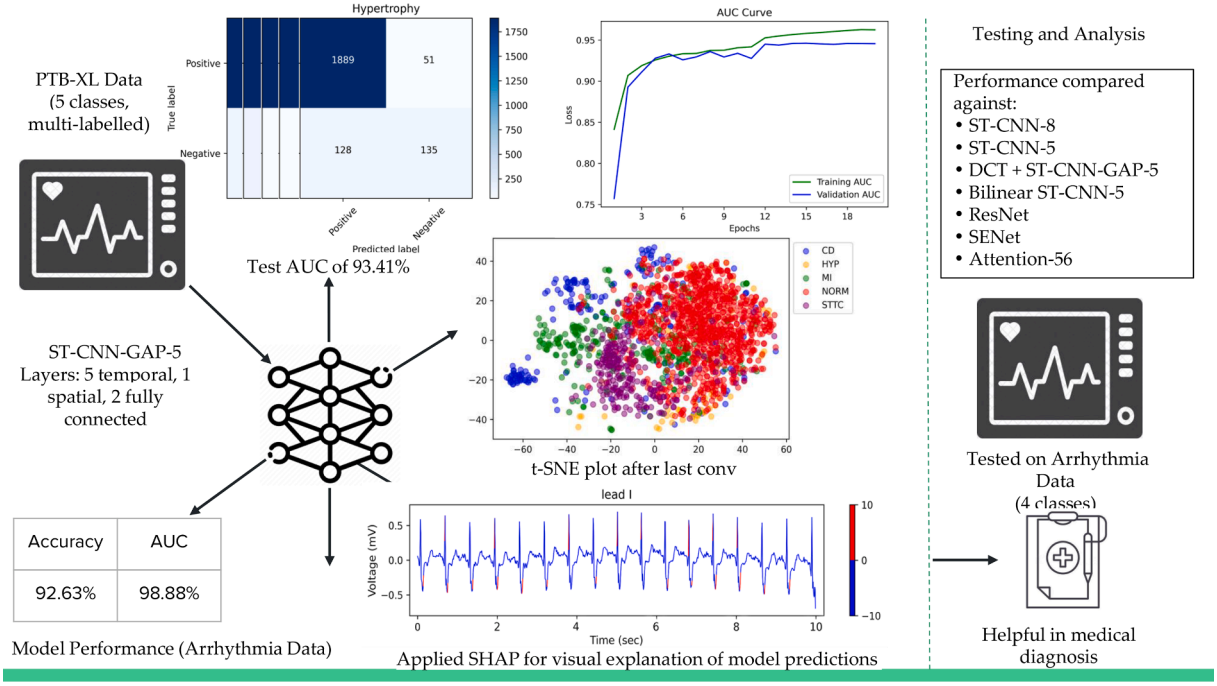


Fig. 2. Work Flow.

Table 3
Class-wise description of PTB-XL dataset (Only major sub-classes with the number of samples (n) are shown for every superclass in column 4).

No. of records	Superclass	Description	Major Subclass (n)
4907	CD	Conduction Disturbance	LAFB (1626), RBBB (1118)
2655	HYP	Hypertrophy	LVH (2137)
5486	MI	Myocardial Infarction	Inferior MI (2685), Anterior MI (2363)
9528	NORM	Normal ECG	Normal ECG (9528)
5250	STTC	ST/T Change	Ischemic (1275), ST changes (770)

testing different state-of-the-art deep learning models. The recent state-of-the-art DL models used for benchmarking in this work and our newly proposed custom-designed DL model are discussed in Section 2.2. Section 2.3 explains the metrics along with averaging techniques that are used to benchmark the performance of different DL models on the problem of cardiac disorder detection in the PTB-XL dataset. Section 2.4 briefly discusses the ShAP (Shapley Additive Explanations) AI interpretability method that has been used to understand the best trained model's decision making and predictions. Section 3 presents the benchmarking results of the proposed DL model against the state-of-the-art methods on both the datasets, followed by interpretability results in Section 3.3. Section 4 discusses the AI interpretability results in the context of actual ECG alterations that happen owing to cardiac disorders and are used by the cardiologists to diagnose a heart disease. Section 5 concludes the paper by summarizing the work done and highlighting its importance towards healthcare. For the ease of readers, list of abbreviations used in this paper are tabulated in Table 2.

2. Materials and Methods

The workflow of this study is presented in Fig. 2. In this study, we have utilized two publicly available datasets that are described next.

2.1. Datasets

PTB-XL, a recently published publicly available ECG dataset available on Physionet [41] is used for benchmarking vaexts models. The ECG data was recorded during 1989–1996 as a part of long term project at the Physikalisch Technische Bundesanstalt. However, the data was made available publicly online only in 2020. It is a 12-lead ECG dataset, and it consists of 21837 samples (each of 10 s length) from 18885 patients, where gender is equally balanced with 52% are male, and 48% are female. Patients age is between 1 to 95 years. It is worth noting that this dataset contains ECGs of various heart pathological conditions including single disease or co-occurrence of many heart diseases in a single patient. This implies that an ECG sample of this dataset may have a single class label (if the subject is suffering with only one disease) or multi-class labels (if the subject is suffering of more than one heart diseases). More importantly, this dataset contains good number of healthy subjects' ECG data as control. This dataset is a very good contribution to the medical and data scientific community.

Two cardiologists labelled 67% of the data and the remaining ECG data was interpreted automatically by an ECG-device, out of which 4.4% of the data was validated by humans. Each ECG data was annotated using some diagnostic statement out of a total of 71 available statements. All these diagnostic statements were further put together into 5 different pathologically relevant classes based on similar pathology. Table 3 describes 5 different diagnostic classes of this dataset and the number of records in each class. It contains ECG waveforms sampled at 500 Hz and 100 Hz. ECG data sampled at 100 Hz is used for all experiments. Some ECG signals were processed to remove spikes at the start and at the end. A measure of the ECG signals' quality is also provided

Table 4
Fold-wise description of PTB-XL dataset.

Fold	Superclass					Count
	CD	HYP	MI	NORM	STTC	
1	481	263	550	941	526	2761
2	487	264	540	967	526	2784
3	487	264	529	993	515	2788
4	494	261	551	928	527	2761
5	496	265	563	941	532	2797
6	496	270	565	932	530	2793
7	479	265	551	970	523	2788
8	492	269	540	935	514	2750
9	497	271	544	957	534	2803
10	498	263	553	964	523	2801
	4907	2655	5486	9528	5250	27826

which includes baseline drift in 7.36% of the signals, static noise and burst noise in 14.94% and 2.81% of signals, respectively.

The dataset is made available in 10 folds for researchers by the authors of the dataset. This implies that the entire dataset is divided into 10 separate partitions/folds. One subject's data occurs in only one of these folds. For our experiments on the PTB-XL dataset, we used the data of the first 9 folds (partitions) for training (88%) and validation (12%), while the data of the last 10th fold was used for testing. Fold-wise description of this dataset excluding the samples with no label/superclass is shown in Table 4. Since some samples have one super-class label and some have more than one label, label-wise description of the dataset is shown in Fig. 3, while the overlap of samples across super-classes is shown in Fig. 4.

We utilized another ECG dataset to further assess the performance of our model and test its generalizability for classifying ECG data. This second dataset is also available publicly and is collected from arrhythmia patients [48]. The dataset was collected by the Chapman University and Shaoxing People's Hospital and contains 10-s duration data of 12-lead ECGs sampled at 500 Hz. The data was collected from 10,646 patients consisting of 5,956 males and 4,690 females majorly (61.1%) lying in the age group of 51–80 years. LOcally WEighted Scatter-plot Smoother (LOESS) and Non Local Means (NLM) were used to remove baseline wander and noise, respectively. It consists of 11 common rhythms that were merged to four superclasses as suggested by Zheng et al. [48]. Total records in each class after removing samples with missing data are shown in Table 5. This dataset (denoised ECG data) was used to ascertain the generalizability of our model for ECG datasets. Of this dataset, 70% records were used for training, 10% for validation and the rest 20% were used as a test set similar to the train-test split by Zheng et al. [48]. Since Zheng et al. [48] did not arrange the data into separate training, validation, and test folds for others to use, it is likely that our test data and their test data may contain different subjects. Similar might be the case of training and validation data on this dataset.

In order to visualize the spread of the data samples in the feature space, t-distributed Stochastic Neighbor Embedding (t-SNE) plots were made of all the ECG samples of both the datasets, considering their entire 10 s long waveforms as the features. For drawing these plots, first the ECG waveforms was averaged across all 12 channels to generate one waveform of 10 s consisting of 1000 samples. Next, t-sne plots are made that are shown in Fig. 5. This figure clearly shows that the ECG samples are mixed across classes in the feature space, i.e, ECG samples of one class does not form a visually distinguishable cluster in the feature space and hence, a classifier is required to learn the distinguishing features to classify the data.

2.2. Models

The following deep learning models were trained and tested on the PTB-XL dataset:

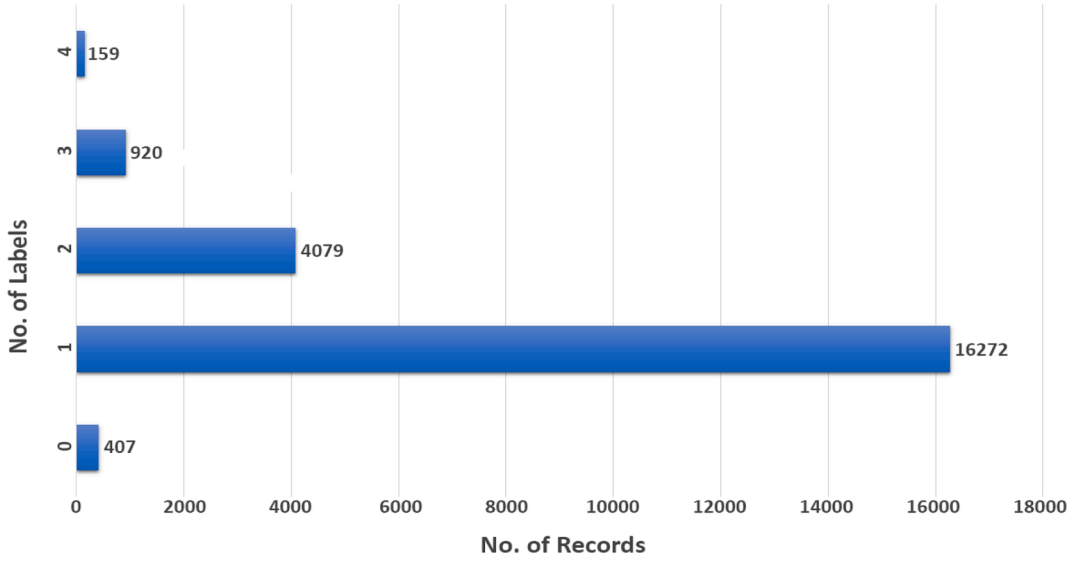


Fig. 3. Label-wise description of PTB-XL dataset (For example, 159 ECG records have 4 class labels and thus, belong to 4 super-classes listed in Table 3.

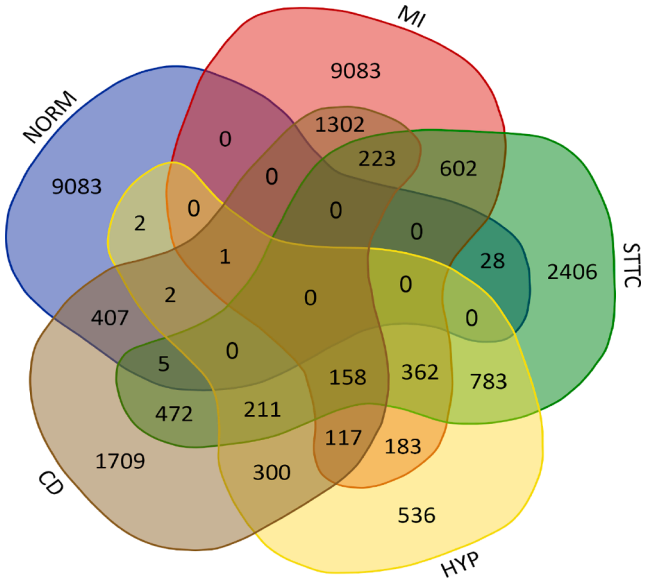


Fig. 4. Venn Diagram for PTB-XL classes.

Table 5
Class-wise description of Arrhythmia dataset.

No. of records	Superclass	Description
2218	AFIB	Atrial Fibrillation
2260	GSVT	Grouped Supraventricular Tachycardia
3888	SB	Sinus Bradycardia
2222	SR	Sinus Rhythm

Spatio-Temporal CNN (ST-CNN-8): Inspired by the work of Attia et al. [6], we considered their CNN model with eight temporal layers (kernel and pool sizes were along a lead), one spatial layer (kernel and pool sizes were across the leads), followed by two dense layers and a final output sigmoid layer as the first model for our experiments. We also labeled this architecture as ST-CNN-8.

ST-CNN-5: To the above specified ST-CNN-8 model, we added two skip connections as shown in Fig. 6 and reduced the temporal layers to five. A lesser number of layers helped in reducing model complexity and the total number of parameters. Its impact on model performance was compensated by using two skip connections covering the five temporal layers.

ST-CNN-GAP-5: Here, in the ST-CNN-5 model, we used global average pooling (GAP) instead of max pool in the last convolution layer. This drastically reduced the total number of trainable parameters from 8,078,309 to 165,061. This architecture is shown in Fig. 6. We call this architecture ST-CNN-GAP-5.

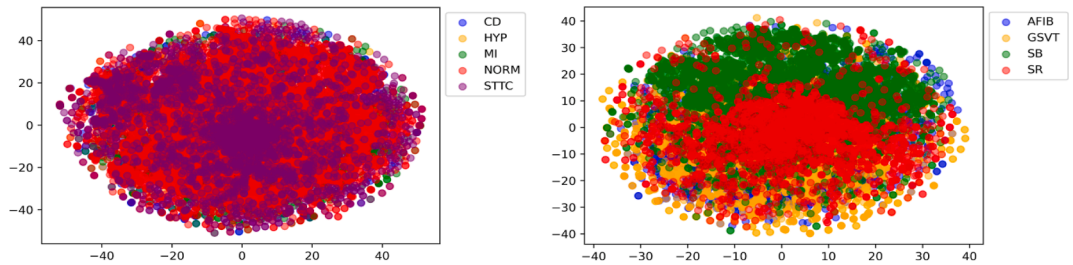
DCT + ST-CNN-GAP-5: Discrete cosine transform (DCT) represents a finite sequence of data into elementary frequency components using cosine functions. DCT is widely used for ECG data compression and feature extraction purposes. ECG classification models trained on DCT extracted features have produced good results and suggested their readiness for clinical settings [27,14]. We applied DCT to the dataset, and its output was used as an input to the ST-CNN-GAP-5 model. In one model, horizontal (along the leads) DCT was applied, while in the other, horizontal DCT followed by vertical (across the leads) was applied. We used type-2 DCT, according to which data points $x_0, x_1, \ldots, x_{N-1}$ are transformed into $X_0, X_1, \ldots, X_{N-1}$ using the given formula:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right] \quad \text{for } k = 0, 1, \ldots, N-1. \tag{1}$$

ResNet: ResNet models with a different number of layers derived from He et al. [18] are used. It uses an improved version of identity mapping in the ResNet, allowing forward and backward signals to directly propagate from one unit to another unit. It makes training easier and generalizes better than the original ResNet. We used the implementation made publicly available by [22].

Bilinear ST-CNN-5: A single CNN is not always efficient in visual recognition. Bilinear CNN, as the name suggests, uses two different CNN models to extract features from the images. The two models can capture different aspects of the images, and later they are merged into a single model. Previous works have suggested that it could produce better results than single CNN models [24]. Performance was measured on Bilinear ST-CNN-5 using three different merge operations: Multiply (element-wise multiplication), Concatenation (concatenate inputs from both models) and Outer Product (multiply each element of one model with each element of the other model).

SENet: Squeeze-and-Excitation Network (SENet) [20] weighs each input channel adaptively depending upon its relevance unlike simple CNN which weights each channel equally. SENet can be easily added to existing architectures and offer performance boost with negligible



(a) PTB-XL (b) Arrhythmia Dataset

Fig. 5. t-SNE Plots of all the ECG samples of both the datasets, considering their entire 10 s (1000 samples) long waveform (averaged across all 12 channels) as the features.

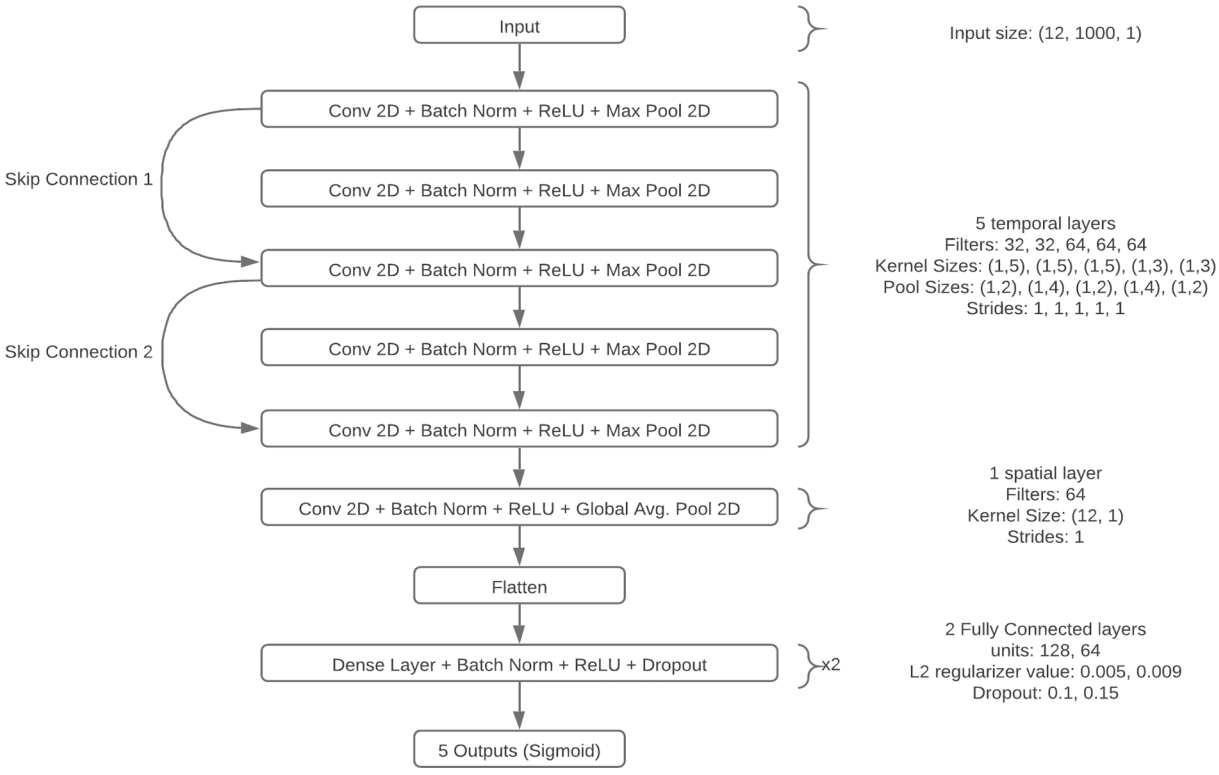


Fig. 6. ST-CNN-GAP-5 Model Architecture.

computational cost. Feature maps are squeezed into a vector of size n, equal to the number of convolutional channels, to obtain a global understanding of each channel. It is then fed into a two-layer neural network, which returns an output vector of the same size. These n values are then used as weights on the original features maps, thus representing each channel based on its importance. We have added SENet block between temporal and spatial layers.

Attention-56: It is based on the residual attention network as described in [42]. Residual attention networks stack attention modules between residual units that can create attention-aware features. It leads to better performance of models on the increased number of layers. We used the implementation made publicly available by [34].

The performance of all the above models has been compared with the state-of-the-art Resnet model named "resnet1d_wang," proposed by [43] and adapted by [35] for one-dimensional inputs. It stacks three residual blocks, followed by a global average pooling layer and an output layer. So far, Strodthoff et al. [35] has reported best results on this multi-class multi-labelled dataset (Table 6).

2.3. Evaluation Metrics

The classification performance on the PTB-XL dataset was computed on all the above discussed state-of-the-art CNN models and our custom-designed models. For comparison, we have used macro-averaged AUC as our primary evaluation metric, which is same for the model of [35] that has so far performed best on this dataset. We have compiled the performance on the following evaluation metrics:

- Binary Accuracy: It is the fraction of correct predictions with respect to the total number of predictions. Here, we have globally computed the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) to find the accuracy.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{2}$$

This is to clarify that binary accuracy has been used on the PTB-XL dataset because this dataset is not only multi-class but also multi-labelled. Since a sample can belong to multiple classes, it implies that it may not be incorrectly predicted in all classes it belongs to,

Table 6
Classification performance of different models.

Model	#params	Train				Test			
		Accuracy	macro AUC	macro AUPRC	micro F1	Accuracy	macro AUC	macro AUPRC	micro F1
resnet1d_wang [35]	-	-	-	-	-	-	93.00	-	-
ST-CNN-8 [6]	15,877,269	92.74	96.87	91.11	85.12	88.92	92.21	80.91	77.28
ST-CNN-5	8,078,309	91.44	96.01	88.64	82.32	88.44	91.86	80.41	76.09
ST-CNN-GAP-5	165,061	**92.41**	**96.72**	**90.55**	**84.67**	**89.73**	**93.41**	**83.39**	**79.28**
DCT (H) + ST-CNN-GAP-5	165,061	87.78	91.67	79.83	74.03	84.42	86.98	70.89	66.59
DCT (HV) + ST-CNN-GAP-5	165,061	88.13	92.14	80.44	74.93	84.44	86.41	70.60	66.96
Bilinear ST-CNN-5 (Concatenate)	16,189,349	91.53	95.96	88.67	82.64	89.51	93.22	83.07	78.58
Bilinear ST-CNN-5 (Multiply)	16,181,157	91.22	95.56	88.12	81.96	88.72	92.29	81.27	76.87
Bilinear ST-CNN-5 (Outer Product)	17,221,541	91.55	96.04	89.00	82.76	89.42	93.20	82.37	78.40
ResNet-18	11,247,877	94.59	98.06	94.28	89.07	87.40	90.55	78.58	74.34
ResNet-34	21,359,749	92.49	96.66	90.80	84.35	88.25	91.74	80.34	75.52
ResNet-50	23,791,877	94.50	97.89	93.75	88.73	87.46	90.73	78.15	74.02
ResNet-101	42,810,117	93.59	97.48	92.60	86.64	88.02	90.50	78.49	74.84
Attention-56	29,809,541	90.47	94.64	86.22	79.86	87.91	91.07	79.33	74.45
SENet	88,046	90.10	94.70	86.14	79.84	88.50	92.24	81.30	76.60

while doing the classification. In other words, let us say, a sample belongs to class 0 and class 2. It may get correctly classified to class 0, but may miss class 2. In this case, we cannot say that the model is entirely wrong. If we frame this problem as binary classification for each class separately, then the correct prediction of class 0 as well as incorrect prediction on class 2 can be captured in this metric. In fact, we are not doing binary classification, but we are working with binary labels for each class. Thus, binary accuracy is used in a way to classify whether a single data item belongs to a specific class (label 1) or not (label 0). This is done for each of the five classes.

- Precision: Precision measures what fraction of predictions are actually correct out of all positive predictions.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

- Recall: Recall measures what fraction of actual positives were correctly identified.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

- F1 score: It is the harmonic mean of precision and recall.

$$F1 - Score = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{5}$$

- AUC (Area Under the ROC Curve): AUC measures class separability by plotting a receiver operating characteristics curve with True Positive Rate (TPR) against the False Positive Rate (FPR) at different classification thresholds, where TPR is on the y-axis and FPR is on the x-axis. An AUC of 0 indicates that all predictions by the model are incorrect. An AUC of 0.5 indicates that the model does not have any class separation ability. An AUC of 1 indicates excellent class separation ability. In macro AUC, AUC is computed separately for each label and then averaged across labels.

$$TPR \quad \left(Recall\right) = \frac{TP}{TP + FN}; \quad FPR = \frac{FP}{TN + FP} \tag{6}$$

- AUPRC (Area under Precision-Recall Curve): AUPRC is a useful metric for imbalanced datasets. AUPRC measures class separability by plotting a precision-recall curve, where Precision is on the y-axis and Recall is on the x-axis. AUPRC has a baseline equal to the fraction of positive data samples. A value greater than the baseline value indicates better class separability.

For evaluating these metrics, we can apply different averaging techniques. These averaging techniques differ in the manner in which they combine results from different classes. In the case of PTB-XL dataset, we have applied them in AUC, AUPRC and F1 score. In the case of Arrhythmia dataset, we have applied them in AUC, AUPRC, Precision, Recall and F1. These techniques are described below:

- Micro-averaging: It calculates metric globally by combining samples across labels. It is preferred in case of class imbalance because it assigns same weight to each instance. This can be understood by recollecting that precision is the ratio of true positive to the sum of true positive and false positive. So, for n classes, our micro-precision would be:

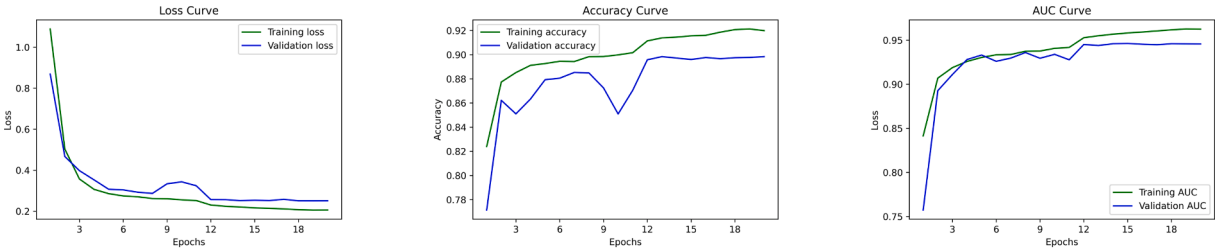
$$P_{micro} = \frac{TP_1 + TP_2 + \ldots + TP_n}{(TP_1 + TP_2 + \ldots + TP_n) + (FP_1 + FP_2 + \ldots + FP_n)} \tag{7}$$

- Macro-averaging: It calculates metric for each class and then finds their average. It is preferred when each class needs to be treated equally. For n classes, our macro-precision with P as precision would be:

$$P_{macro} = \frac{P_1 + P_2 + \ldots + P_n}{n} \tag{8}$$

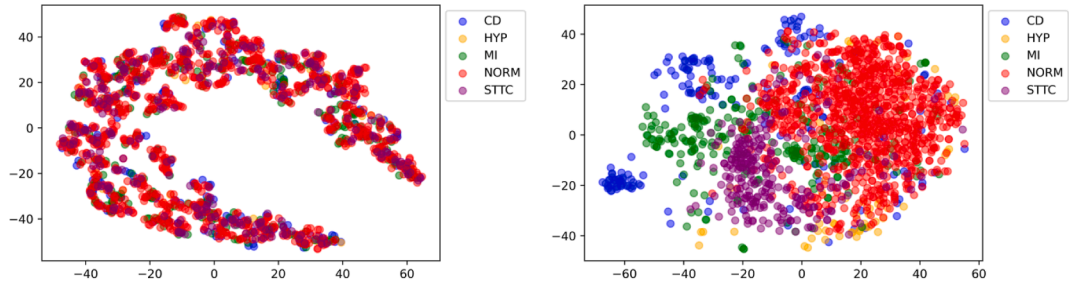
- Weighted-averaging: It calculates metric for each class and then finds their weighted average. Here, weights are defined as the number of samples for each class out of all samples. For n classes, our weighted-precision with w as weight and P as precision would be:

$$P_{weighted} = \frac{w_1 P_1 + w_2 P_2 + \ldots + w_n P_n}{n} \tag{9}$$



(a) Loss Curve (b) Accuracy Curve (c) AUC Curve

Fig. 7. Curves for ST-CNN-GAP-5.



(a) First Conv Layer (b) Last Conv Layer

Fig. 8. t-SNE Plots after first and last conv layers on test set of PTB-XL Dataset.

2.4. AI Interpretability Method

Although CNN models have demonstrated huge success in many tasks, they are considered as "black boxes" due to the lack of interpretability of their internal working. SHAP (Shapley Additive Explanations) developed by Lundberg et al. [25] can help in providing visual explanations or interpretability of the predictions made by machine learning models. It assigns importance value to each feature based on its impact on the output prediction. We used SHAP Gradient Explainer to explain predictions of our best trained model, ST-CNN-GAP-5, on each lead of the ECG data of a patient. It is based on integrated gradients method proposed by Sundararajan et al. [36], which is a feature attribution method for deep neural networks. It approximates a SHAP value for each input feature. From these SHAP values, we used the top 500 SHAP values as the important features that contribute to the diagnosis of a particular ECG record.

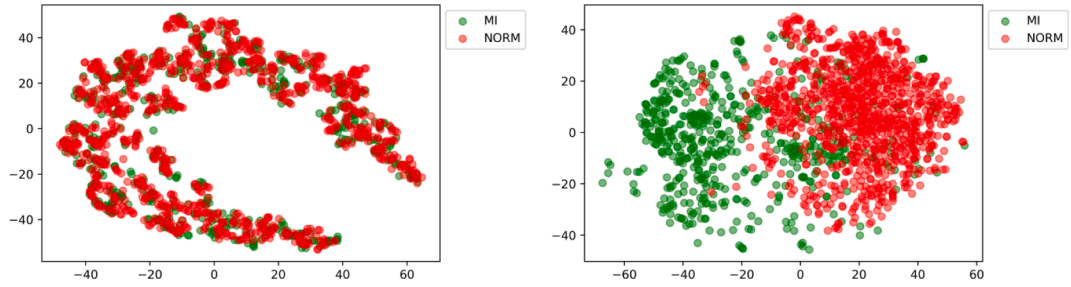
3. Results and Analysis

All the models were trained on the PTB-XL dataset that was divided into training and test using the folds as recommended by Wagner et al. [41]. The first 9 folds were used for training and validation, while the last fold was used for testing. It was made sure that no test set sample was shown during the training phase.

All the models were trained on Google Colab, a cloud-based Jupyter notebook environment. GPU was used as a hardware accelerator. Keras API, which runs on top of the Tensorflow framework, was used to implement the models. Scikit-learn was used to compute different metrics. To train the multi-labelled dataset, we used sigmoid cross-entropy for performing binary classification for each class. Adam optimization algorithm was used with a learning rate of 0.0005 and beta value of 0.9.

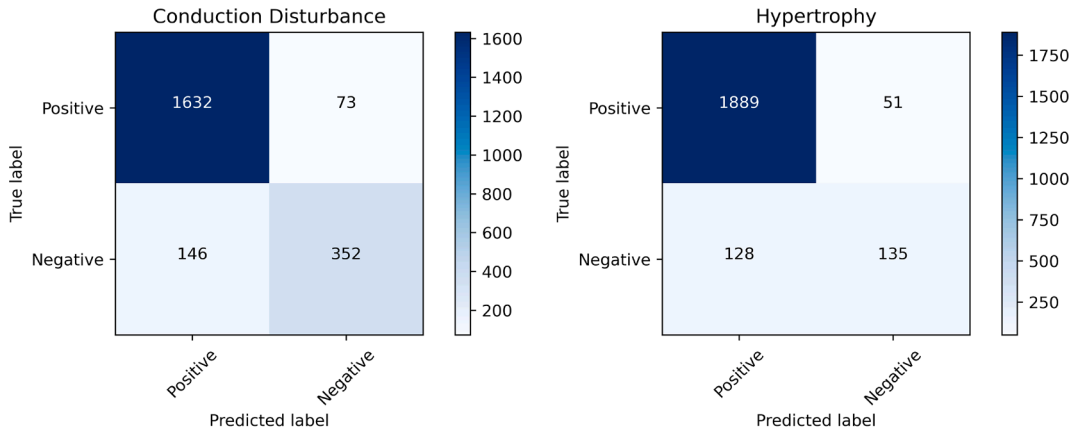
3.1. Performance on PTB-XL Dataset

The experiment results on all models are shown in Table 6. The table contains the total number of trainable parameters, accuracy, macro AUC, macro AUPRC and micro F1 score on both train and test set for each model on the multi-labelled dataset. It includes the results on the state-of-the-art method [35] along with results from our models. Results show that the *ST-CNN-GAP-5* model yields the best performance. The loss curve, accuracy curve and AUC for curve for best performing model is shown in Fig. 7. To visually demonstrate the ability of ST-CNN-GAP-5 to distinguish between different classes, we have added t-SNE plots of the output of first and last convolution layers in Fig. 8. This figure is shown for test records that have single class-label. This graph clearly



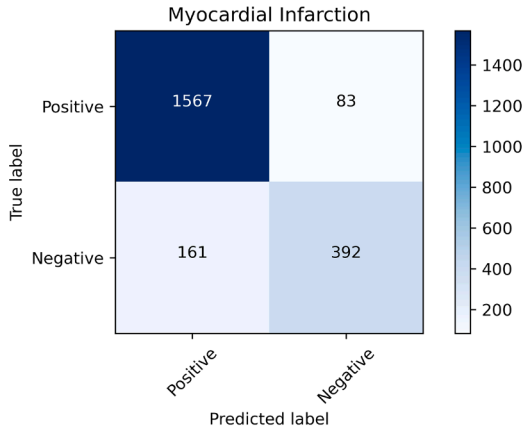
(a) First Conv Layer (b) Last Conv Layer

Fig. 9. t-SNE Plots after first and last conv layers on test set of MI-NORM Samples.

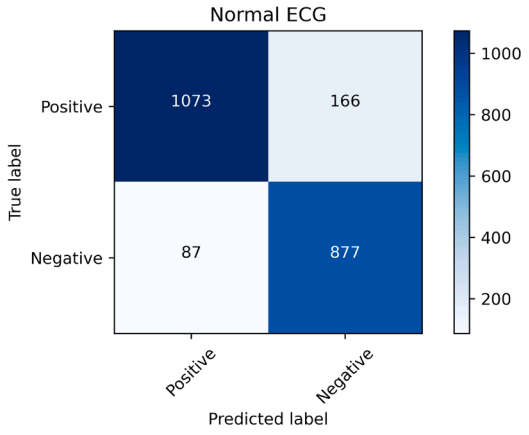


(a) Conduction Disturbance

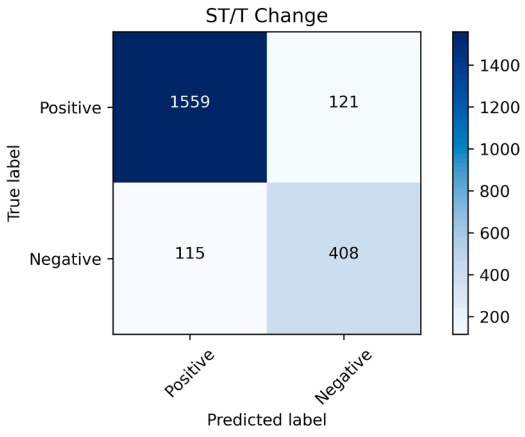
(b) Hypertrophy



(c) Myocardial Infarction



(d) Normal ECG



(e) ST/T Change

Fig. 10. Confusion Matrices for ST-CNN-GAP-5 on test set.

demonstrates that samples from different classes started forming clusters as we moves from the first convolution layer to the last convolution layer. The utility of this architecture is more clearly visible on the subset of samples containing only two classes, MI and NORM (Fig. 9). From these two figures, we can infer that our proposed model learned the distinguishing features and gained the ability to separate samples from different classes. The confusion matrices for ST-CNN-GAP-5 on the test set are shown in Fig. 10.

The ST-CNN-8 model yielded an AUC of 92.21%. ST-CNN-5 demonstrated slightly lesser performance over the ST-CNN-8 model. Standard ResNet models performed poorly compared to even the ST-CNN-5 model and its variants. AUC increased with increasing the number of layers from 18 to 34, but it dipped on further increasing the number of layers in the ResNet models. Attention-56 model scored an AUC of 91.07%. SENet, with the least number of trainable parameters, scored 88.50% accuracy and 92.24% AUC. Bilinear ST-CNN-5 models performed comparatively better, with 93.22% and 93.20% achieved by concatenation and outer product operations, respectively. However, discrete cosine transformation resulted in the lowest performance among all models, possibly because the input images did not have much

Table 7
Classification performance on the Test set of Arrhythmia Database.

Metric	Zheng et al. 2020 500 Hz	Yildirim et al. 2020 500 Hz	Rieg et al. 2020 500 Hz	Proposed ST-CNN-GAP-5 500 Hz	Proposed ST-CNN-GAP-5 100 Hz
Accuracy	–	96.13	93.67	95.85	96.22
Macro AUC	–	–	–	99.46	99.54
Macro AUPRC	–	–	–	98.53	98.73
Macro Precision	96.60	95.78	92.60	95.44	95.90
Weighted Precision	97.10	–	–	95.85	96.25
Macro Recall	96.40	95.43	92.73	95.34	95.71
Weighted Recall	97.00	–	–	95.85	96.22
Macro F1	96.50	95.57	92.66	95.39	95.79
Weighted F1	97.00	–	–	95.84	96.22

Table 8
Hyperparameters for fine-tuned ST-CNN-GAP-5 on Arrhythmia Dataset.

Layers	Data Frequency	500 Hz	100 Hz
Input Layer	Size	(12, 5000, 1)	(12, 1000, 1)
5 Temporal Layers	Filters	32, 32, 32, 64, 64	32, 32, 32, 64, 64
	Kernel Sizes	(1,55), (1,55), (1,55), (1,53), (1,53)	(1,15), (1,15), (1,15), (1,13), (1,13)
	Pool Sizes	(1,2), (1,4), (1,2), (1,4), (1,2)	(1,2), (1,4), (1,2), (1,4), (1,2)
	Strides	1, 1, 1, 1, 1	1, 1, 1, 1, 1
1 Spatial Layer	Filters	64	64
	Kernel Sizes	(12,1)	(12,1)
	Strides	1	1
2 Fully Connected Layers	Units	128, 32	128, 64
	L2 regularizer values	0.01, 0.02	0.009, 0.015
	Dropout	0.2, 0.25	0.2, 0.25

discrimination in the DCT domain. ST-CNN-GAP-5 achieved the best performance with 89.73% accuracy, AUC score of 93.41%, AUPRC score of 83.39% and F1 score of 79.28%. Global Average Pooling (GAP), which outputs the average of each feature map, reduced the vector size to 64 before the dense layers. It also helped reduce the total trainable parameters to 165,061, which is much lower compared to other models. This reduced overfitting and could be the reason for the better performance.

Three models, Bilinear ST-CNN-5 (Concatenate), Bilinear ST-CNN-5 (Outer Product) and ST-CNN-GAP-5, were able to perform better than the state-of-the-art results. They yielded an AUC of 93.22%, 93.20%, and 93.41%, which is an improvement of 0.22%, 0.20 and 0.41%, respectively, over the previous state-of-the-art AUC of 93.00%. ECG is time-series data representing electrical impulses. In the case of image data, we used rectangular filters that may not work very well in ECG data. The reason behind the improvement in these models could be the use of spatial and temporal layers that could exploit the temporal information of all the channels as well as the information present across channels.

3.2. Performance on Arrhythmia Dataset

Generalization of deep learning models has always remained a challenge for researchers [28]. Models tend to work exceptionally well on one dataset, but fail to replicate the same results on other similar dataset. Hence, it is important to ascertain whether a model is actually solving a problem or is only giving good results on one particular dataset. We used the network architecture of ST-CNN-GAP-5 with hyperparameter tuning to train another model on arrhythmia dataset of over 10,000 patients. The hyperparameters for the fine-tuned ST-CNN-GAP-5 are presented in Table 8. Zheng et al. [48] generated this arrhythmia dataset and could achieve an average F1-score of 97% on the test data using an extreme gradient tree classification model. They used the dataset at 500 Hz sampling ratio. They used 230 features extracted from the ECG leads along with age and gender to boost the performance of their model. Similarly, Yildirim et al. and Rieg et al. [47,30] utilized the dataset at 500 Hz. Yildirim et al. [47] developed a DNN model which includes both representation learning and sequence learning tasks. They have used single ECG leads as input and obtained the best overall performance from lead II. Rieg et al. [30] built a tree-based model using 24 features extracted from ECG signals, of which 13 features are same as that used by Zheng et al. [48]. Their performance on the test data along with the performance results from our model are shown in Table 7. We generated results on both 500 Hz dataset as well as at a downsampled version of 100 Hz. From Table 7, it is observed that the proposed ST-CNN-GAP-5 performed comparable on all the metrics. Further, the

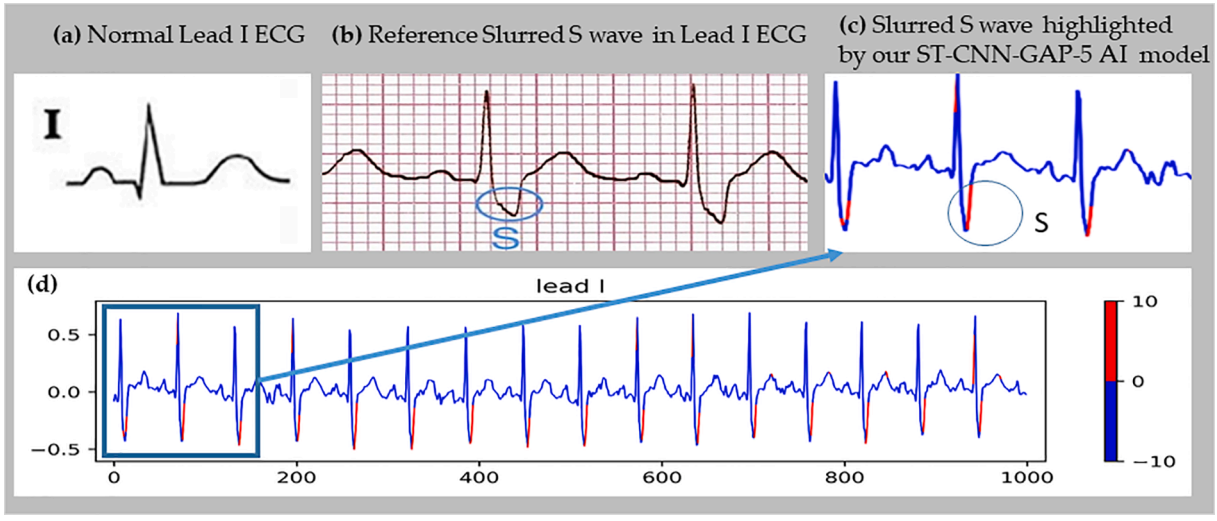


Fig. 11. Slurred S wave highlighted in Lead I of ECG of a patient having Right Bundle Branch Block– 11(a): Showing Lead I of a normal ECG; 11(b): slurred S wave in lead I was observed in previous literature and studies; and 11(c): a part of ECG (shown at the bottom of the figure) pointing to slurred S wave highlighted in red color by our AI model. (On Y axis- left side each unit is equal to 10 mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

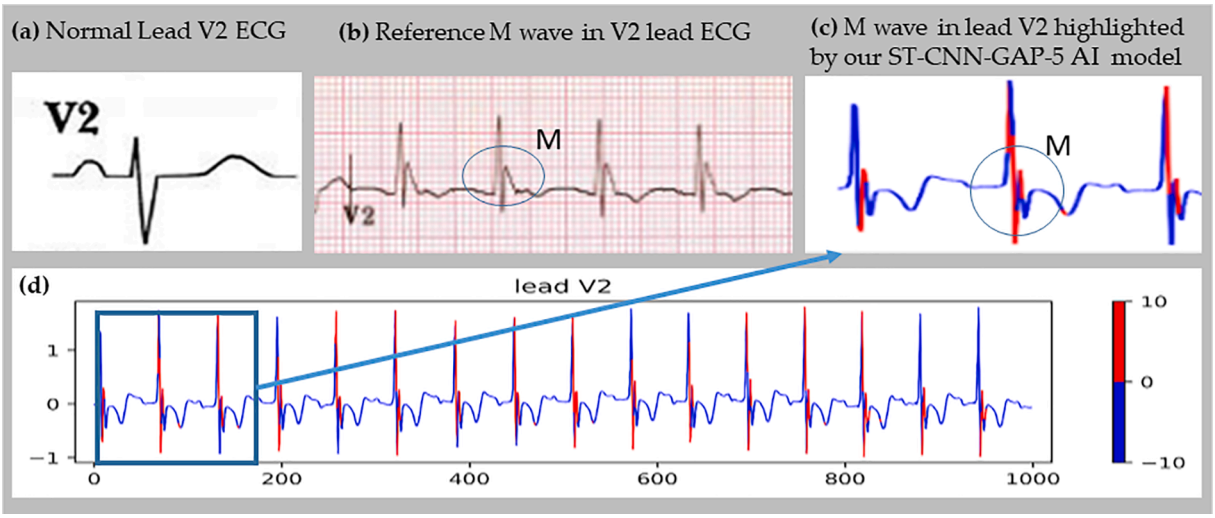


Fig. 12. M wave highlighted in Lead V2 of ECG of a patient having Right Bundle Branch Block– 12(a): Showing Lead V2 of a normal ECG; 12(b): M wave in lead V2 was observed in previous literature and studies; and 12(c): a part of ECG (shown at the bottom of the figure) pointing to M wave highlighted in red color by our AI model. (On Y axis- left side each unit is equal to 10 mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

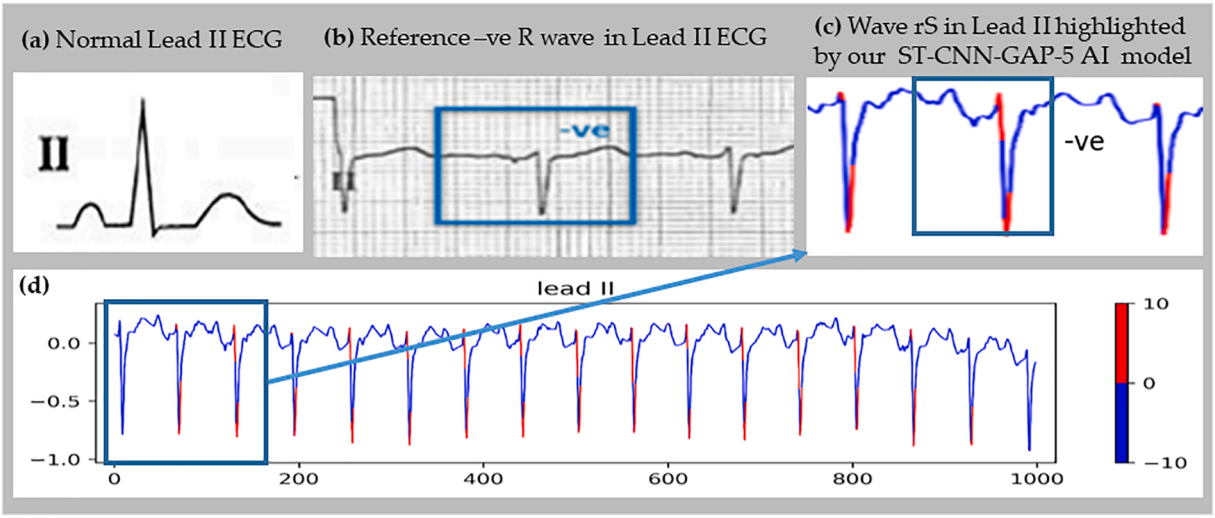


Fig. 13. Negative R wave (rS wave) highlighted in Lead II of ECG of a patient having Left Anterior Fascicular Block– 13(a): Showing Lead II of a normal ECG; 13(b): negative R wave (rS wave) in lead II was observed in previous literature and studies; and 13(c): a part of ECG (shown at the bottom of the figure) pointing to negative R wave (rS wave) highlighted in red color by our AI model. (On Y axis- left side each unit is equal to 10 mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

performance at 100 Hz is better than that at 500 Hz. This is because the ECG data is generally rich in low frequency and hence, 100 Hz sampling ratio is sufficient to capture the features. perhaps, 500 Hz signal is adding a lot of redundant features and hence, the performance is slightly inferior compared to 100 Hz. Further, all the metrics crossed 95% at both 500 Hz and 100 Hz. This shows that the proposed model could separate samples of distinct classes.

3.3. *Interpretability of the best trained DL Model*

Since our proposed ST-CNN-GAP-5 model performed best on the PTB-XL dataset, we used SHAP (Shapley Additive Explanations) interpretability on this model trained on the PTB-XL dataset. This dataset consists of five superclasses. Each superclass has a number of subclasses, of which only the major contributing subclasses are shown in Table 3. In this subsection, we have demonstrated the interpretability of DL model on each of these major subclasses.

The ECG wave/segments, highlighted in red color, are based on the top 500 SHAP values. These are the most important features for the DL classification, while the features with lesser importance are seen in the blue color. Our model on PTB-XL dataset identified the slurred S wave in Lead I of Fig. 11 and M wave in Lead V2 of Fig. 12 indicating the Right Bundle Branch Block (RBBB disorder that belongs to the superclass Conduction Disturbance). Similarly, rS complex in Leads II, III, and aVF in Fig. 13–15 got highlighted in red. These highlighted portions are indeed the indicators of Left Anterior Fascicular Block (LAFB disorder that belongs to the superclass Conduction Disturbance) and the findings are consistent with the diagnostic criteria of conduction disturbances. Prominent Q wave in Leads III and aVF indicates inferior wall myocardial infarction as rightly observed in Fig. 16. On the other hand, a prominent Q wave in Leads V2 and V4 indicates anteroseptal myocardial infarction (belong to major superclass Myocardial Infarction) as captured in Fig. 17. A sum of R wave in Lead I and S wave in Lead III greater than 25 mm is characteristic of Left ventricular hypertrophy

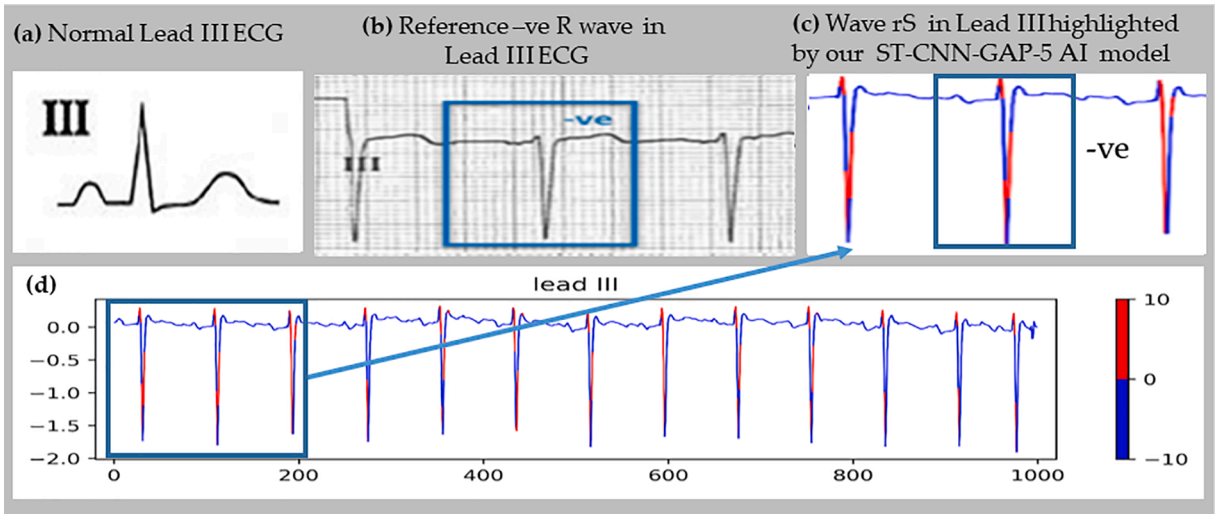


Fig. 14. Negative R wave (rS wave) highlighted in Lead III of ECG of a patient having Left Anterior Fascicular Block– 14(a): Showing Lead III of a normal ECG; 14(b): negative R wave (rS wave) in lead III was observed in previous literature and studies; and 14(c): a part of ECG (shown at the bottom of the figure) pointing to negative R wave (rS wave) highlighted in red color by our AI model. (On Y axis- left side each unit is equal to 10 mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

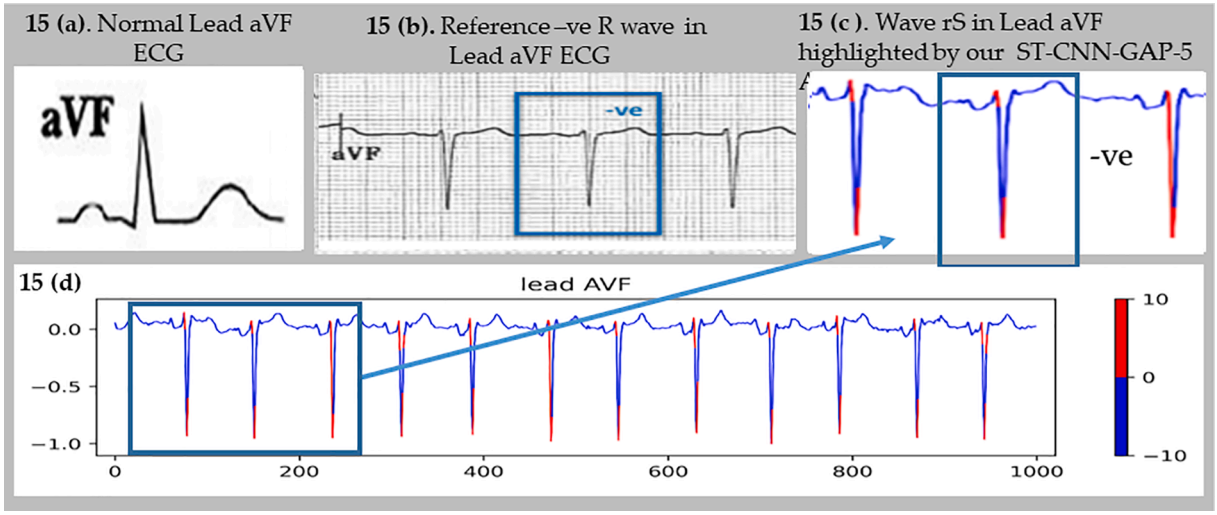


Fig. 15. Negative R wave (rS wave) highlighted in Lead aVF of ECG of a patient having Left Anterior Fascicular Block– 15(a): Showing Lead aVF of a normal ECG; 15 (b): negative R wave (rS wave) in lead aVF was observed in previous literature and studies; and 15(c): a part of ECG (shown at the bottom of the figure) pointing to negative R wave (rS wave) highlighted in red color by our AI model. (On Y axis- left side each unit is equal to 10 mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

(LVH, that belongs to the superclass hyperthrophy) that is captured in a patient's ECG as shown in Fig. 18. Likewise, LVH is also represented by a sum greater than 35 mm of the amplitude of S wave in Lead V1 and the R wave in Leads V5 or V6. This has been highlighted in Fig. 19 of the ECG of a patient. These finding are characteristic features of LVH.

4. Discussion

In this study, a few of our deep learning ECG models performed very well on multi-class heart disease classification. It includes bilinear models and ST-CNN-GAP-5, of which the latter performed best. Such models can also help non-cardiologists in easy diagnosis and triage of patients with chest pain and other symptoms as a screening systems for cardiovascular disorders. Secondly, our model exhibits the explainability/interpretability of the disease class prediction on the ECG waveforms that are characteristic of those cardiac diseases, helping medical doctors and caregivers trust the decisions made by the ML model. This interpretability also works as the validation on the efficacy of the model proposed and trained in this work.

Conduction disorders are the problems associated with the electrical system that is responsible for the control of heart rate and rhythm. Impairment of conduction in either the right or left bundle system leads to Right/Left bundle branch block. ECG changes in RBBB includes QRS duration of more than 120 ms. In Lead I, the duration of S wave is more than that of the R wave. These features are visualized in the shapley plot (Fig. 11). In Lead V2, there is RSR' wave complex, called as "M" wave, where the prime-R (R') is the second positive abnormal R wave and is usually wider than the initial R wave [23], this is clearly observed in Fig. 12. ECG characteristic of Left anterior fasicular block are negative QRS complex in leads II, III, and aVF septal, and rS pattern (small r, deep S) in the inferior leads II, III, and aVF [16], which can be visualized in SHAP plot in Fig. 13–15.

ECG is the mainstay for diagnosis of myocardial infarction. ECG findings depend on various factors including duration (acute vs

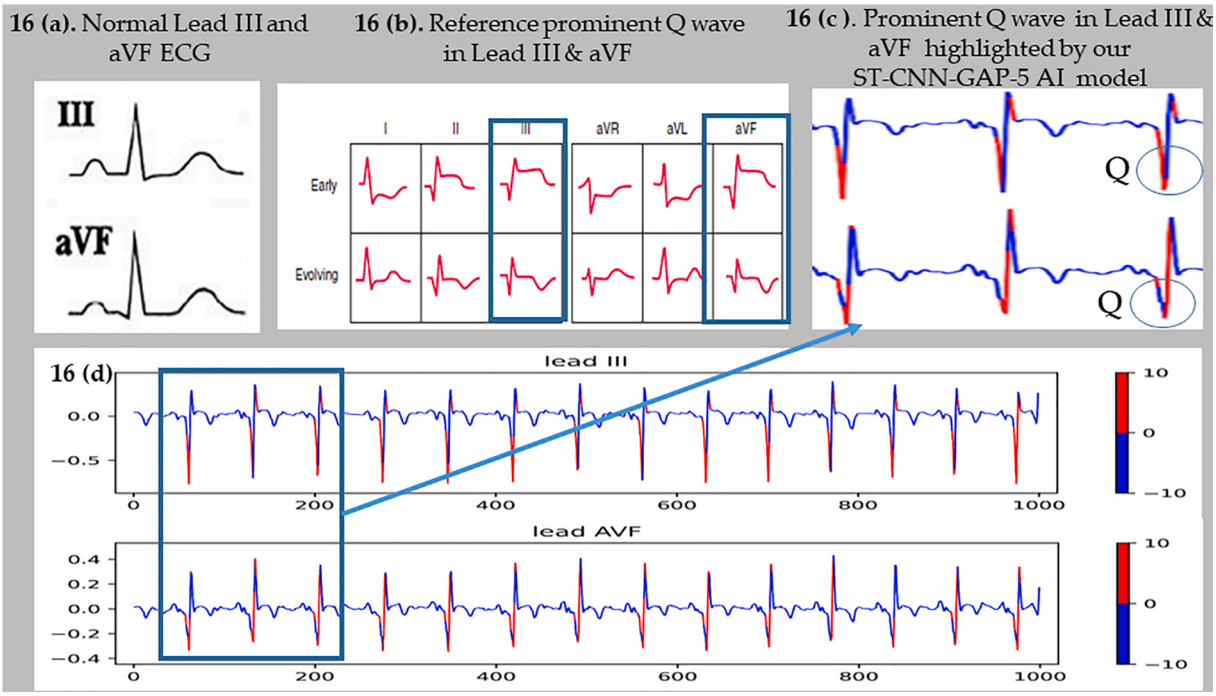


Fig. 16. Prominent Q wave highlighted in Leads III and aVF of ECG of a patient having Inferior myocardial infarction– 16(a): Showing Leads III & aVF of a normal ECG; 16(b): Prominent Q wave observed in previous literature and studies; and 16(c): a part of ECG (shown at the bottom of the figure) pointing to Prominent Q wave highlighted in red color by our AI model. (On Y axis- left side each unit is equal to 10 mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

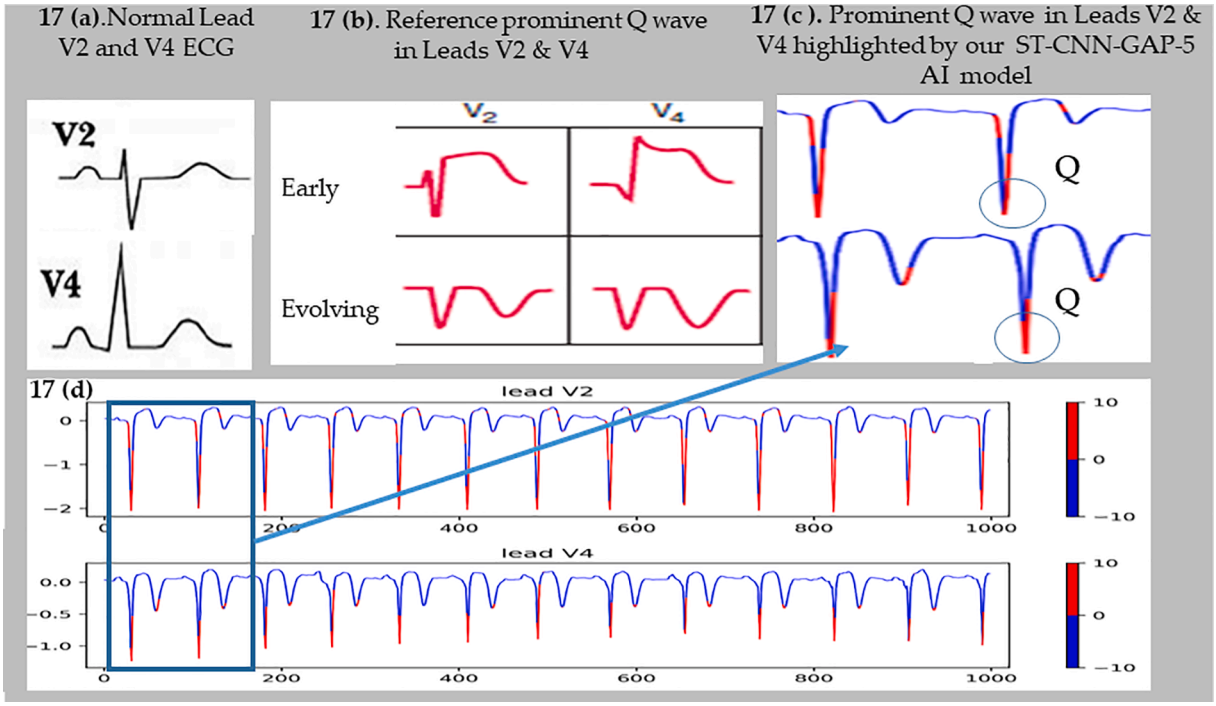


Fig. 17. Prominent Q wave highlighted in Leads V2 and V4 of ECG of a patient having Anteroseptal Myocardial Infarction– 17(a): Showing Leads V2 & V4 of a normal ECG; 17(b): Prominent Q wave observed in previous literature and studies; and 17(c): a part of ECG (shown at the bottom of the figure) pointing to Prominent Q wave highlighted in red color by our AI model. (On Y axis- left side each unit is equal to 10 mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

chronic), localization of MI (anterior, inferior or posterior), and reversible or irreversible ischemia. ST segment elevations occur as the earliest sign and persists for hours. They are typically followed by evolving T-wave inversions and prominent Q waves [40,13]. In our subject with inferior wall MI, the patient had prominent Q waves in leads III and aVF that are clearly visualized in shapley Fig. 16. Similarly, in the ECG waveform of the patient with anteroseptal myocardial infarction, prominent Q waves and T-wave inversions are highlighted in

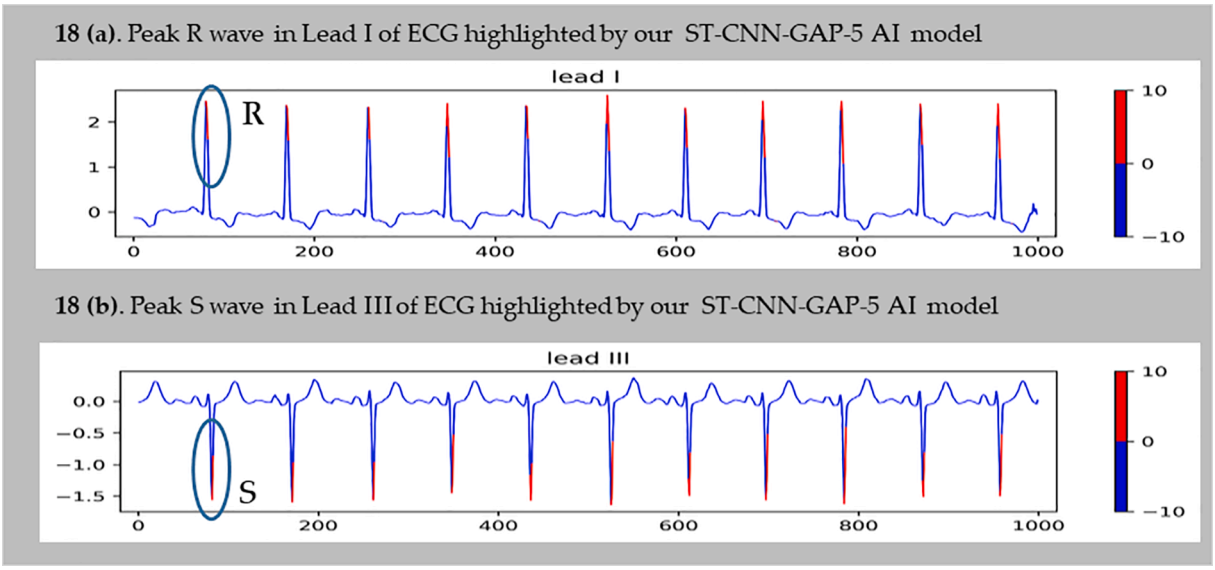


Fig. 18. ECG of a patient having Left Ventricular Hypertrophy: Sum of R wave in lead I and S wave in lead III > 25 mm { 18(a): Highlighted Peak R wave in lead I of ECG, and 18(b): Highlighted Peak S wave in lead III of ECG (On Y axis- left side each unit is equal to 10mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

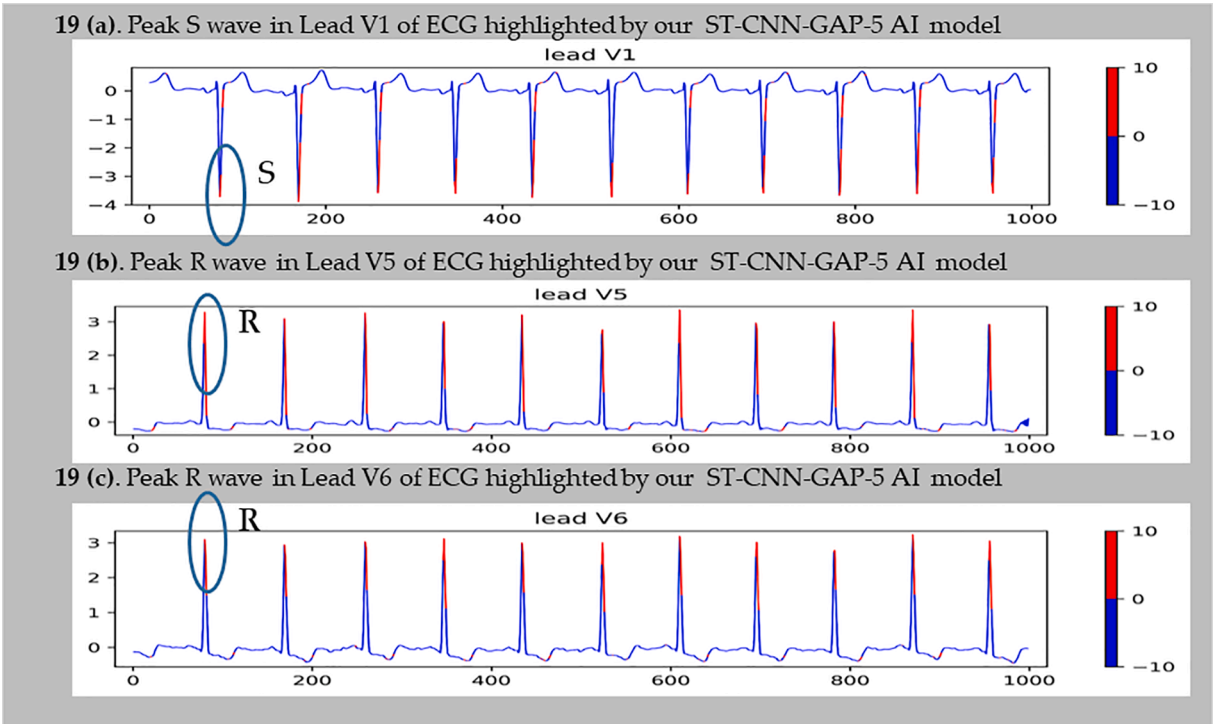


Fig. 19. ECG of a Left Ventricular Hypertrophy: Sum of S wave in V1 and the R wave in V5 or V6 is greater than 35 mm– 19(a): Highlighted Peak S wave in Lead V1 of ECG, 19(b): Highlighted Peak R wave in Lead V5 of ECG, and 19(c): Highlighted Peak R wave in lead V6 of ECG. (On Y axis- left side each unit is equal to 10 mm, right side color bar indicates SHAP values, red color indicated part of ECG highly contribute to AI decision, X-axis is in milliseconds).

V2-V4 as shown in Fig. 17. Left ventricular hypertrophy is a thickening (hypertrophy) of main pumping chamber walls of the heart (left ventricle). The 12-lead ECG is the most common and screening test to determine the presence of LVH in hypertensive patients [45]. To diagnose LVH via ECG, multiple criteria are suggested by Tavares et al. [38] guidelines. One such criteria is Sokolow-Lyon Criteria that states that the sum of S wave in V1 plus the R wave in V5 or V6 is greater than 35 mm. Interestingly, this is clearly evident in the results of our model in Fig. 19, wherein the peak of S and R wave can be visualized clearly. Another criteria is that the sum of R wave in lead I and S wave in lead III is more than 25 mm. This criteria is also met on the data as observed through our trained ML model as evident from the highlighted red colored R and S waves by SHAP in LVH patient's ECG data in Fig. 18, when applied on our trained ML model. Third criteria is that the R-wave in V5 or V6 should be more than or equal to 30 mm. It can be seen clearly from the SHAP applied on our trained model because R-wave in V5 and V6 > 30 mm are highlighted in red color (Fig. 19).

Interpretability of the automated ECG model augments the

effectiveness of cardiologists in diagnosing heart disease accurately with less human error, especially, in overloaded healthcare setup in low/middle-income countries such as India. Furthermore, it helps to gain the confidence and improve the trust of the cardiologist towards an automated ECG AI model. Therefore, it can be implemented at a clinical setup, where low/middle-income countries struggle with a high burden of heart disease and poor healthcare delivery infrastructure. Moreover, explainable AI model on ECG can help the non-cardiologist to diagnose faster without wasting their precious time in ascertaining many heart diseases. Since AI-ECG model augments the clinical decision-making system, this can improve the efficiency of primary and secondary healthcare services.

In this study, we have made a number of contributions.

- We have benchmarked the performance of the recent DL architectures for the detection of cardiac disorders on a large ECG dataset, PTB-XL, that is publicly available.
- We have proposed our custom designed CNN architecture (DL model) and compared its performance with these state-of-the-art methods.
- In order to test the generalizability of our proposed best performing DL model, we have assessed its performance on another ECG dataset of arrhythmia patients.
- Lastly and most importantly, we have visualized and interpreted the portions of ECG waveforms that are used by our DL model in diagnosing heart disorders of PTB-XL dataset and got these validated through clinicians to assess the explainability or interpretability of the decisions of the proposed deep convolution network model for diagnostic purposes.

5. Conclusion

In this paper, we presented results from various models trained using the ECG dataset. It is evident that different variants of Spatio-temporal CNN (ST-CNN-8), including ResNet, SENet etc., worked better than a simple CNN architecture alone. This observation is aligned with the inferences drawn from previous work on the two publicly available ECG datasets. Our ST-CNN-GAP-5 model with skip connections, reduced layers, and reduced parameters via GAP produced better results than the existing state-of-the-art results on these datasets. The proposed ST-CNN-GAP-5 model demonstrated generalizability by giving good results on two different ECG dataset. Interestingly and most importantly, the interpretation of the DL model by SHAP in multiple heart disorders highlighted the same segments of ECG waves as would have been analyzed by the expert cardiologist, while inferring/diagnosing a heart patient. This shows that our proposed model can be easily integrated with the existing ECG machines, to help the doctors in primary and secondary healthcare centres to diagnose faster, accurately and with the proof, so that patient can be timely referred to cardiology centers for further specific treatment. Deployment of such models can help duty doctors in primary and secondary healthcare centres, where cardiologists may not be generally available.

Code Availability

All our trained models along with a few sample files are available at GitHub (https://github.com/tusharkadian/BSPC). We have also created a web application for viewing SHAP interpretability on ECG waveforms (https://ecgdetect.sbilab.iiitd.edu.in).

CRediT authorship contribution statement

Atul Anand: Conceptualization, Methodology, Validation, Software, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Tushar Kadian:** Conceptualization, Methodology, Validation, Software, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Manu Kumar Shetty:** Conceptualization, Methodology, Validation, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Supervision, Resources, Project administration. **Anubha Gupta:** Conceptualization, Methodology, Validation, Investigation, Formal analysis, Writing - original draft, Writing - review & editing, Supervision, Resources, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to express our special thanks to Dikshant Sagar, Research Associate, SBILab, IIIT-Delhi for developing the web-app for this project.

References

[1] S. Agrawal, A. Gupta, Fractal and EMD based removal of baseline wander and powerline interference from ECG signals, Computers in biology and medicine 43 (2013) 1889–1899.
[2] S. Agrawal, A. Gupta, Removal of baseline wander in ECG using the statistical properties of fractional Brownian motion, in: 2013 IEEE International Conference on Electronics, Computing and Communication Technologies, IEEE, 2013, pp. 1–6.
[3] J.P. Allam, S. Samantray, S. Ari, Spec: A system for patient specific ecg beat classification using deep residual network, Biocybernetics and Biomedical Engineering 40 (2020) 1446–1457.
[4] N. Ansari, A. Gupta, WNC-ECGlet: Weighted non-convex minimization based reconstruction of compressively transmitted ECG using ECGlet, Biomedical Signal Processing and Control 49 (2019) 1–13.
[5] M. Arif, I.A. Malagore, F.A. Afsar, Detection and localization of myocardial infarction using k-nearest neighbor classifier, Journal of medical systems 36 (2012) 279–289.
[6] Attia, Z.I., Friedman, P.A., Noseworthy, P.A., Lopez-Jimenez, F., Ladewig, D.J., Satam, G., Pellikka, P.A., Munger, T.M., Asirvatham, S.J., Scott, C.G. et al. (2019a). Age and sex estimation using artificial intelligence from standard 12-lead ECGs. Circulation: Arrhythmia and Electrophysiology, 12, e007284.
[7] Z.I. Attia, P.A. Noseworthy, F. Lopez-Jimenez, S.J. Asirvatham, A.J. Deshmukh, B. J. Gersh, R.E. Carter, X. Yao, A.A. Rabinstein, B.J. Erickson, et al., An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction, The Lancet 394 (2019) 861–867.
[8] F.X. Barthelemy, J. Segard, P. Fradin, N. Hourdin, E. Batard, P. Pottier, G. Potel, E. Montassier, ECG interpretation in emergency department residents: an update and e-learning as a resource to improve skills, European Journal of Emergency Medicine 24 (2017) 149–156.
[9] Y. Birnbaum, J.M. Wilson, M. Fiol, A.B. de Luna, M. Eskola, K. Nikus, ECG diagnosis and classification of acute coronary syndromes, Annals of Noninvasive Electrocardiology 19 (2014) 4–14.
[10] A. Boehm, X. Yu, W. Neu, S. Leonhardt, D. Teichmann, A novel 12-lead ECG T-shirt with active electrodes, Electronics 5 (2016) 75.
[11] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (2019) 832.
[12] Y.-J. Chen, C.-L. Liu, V.S. Tseng, Y.-F. Hu, S.-A. Chen, Large-scale classification of 12-lead ECG with deep learning, in: 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), IEEE, 2019, pp. 1–4.
[13] J.-P. Collet, H. Thiele, E. Barbato, O. Barthélémy, J. Bauersachs, D.L. Bhatt, P. Dendale, M. Dorobantu, T. Edvardsen, T. Folliguet, et al., 2020 ESC Guidelines for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation: the Task Force for the management of acute coronary syndromes in patients presenting without persistent ST-segment elevation of the European Society of Cardiology (ESC), European Heart Journal 42 (2021) 1289–1367.
[14] U. Desai, R.J. Martis, C.G. Nayak, K. Sarika, S.G. Nayak, A. Shirva, V. Nayak, S. Mudassir, Discrete cosine transform features in automated classification of cardiac arrhythmia beats, in: Emerging research in computing, information, communication and applications, Springer, 2015, pp. 153–162.
[15] Diker, A., Cömert, Z., Avcı, E., Toğaçar, M., & Ergen, B. (2019). A novel application based on spectrogram and convolutional neural network for ECG classification. In: 2019 1st International Informatics and Software Engineering Conference (UBMYK) (pp. 1–6). IEEE.
[16] M.-Y. Gao, Y. Tian, L. Shi, Y.-J. Wang, B.-Q. Xie, J. Qi, L.-J. Zeng, X.-X. Li, X.-C. Yang, X.-P. Liu, Electrocardiographic morphology during left bundle branch area pacing: characteristics, underlying mechanisms, and clinical implications, Pacing and Clinical Electrophysiology 43 (2020) 297–307.

[17] 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.
[18] K. He, X. Zhang, S. Ren, J. Sun, Identity mappings in deep residual networks, in: European conference on computer vision, Springer, 2016, pp. 630–645.
[19] S.A. Hicks, J.L. Isaksen, V. Thambawita, J. Ghouse, G. Ahlberg, A. Linneberg, N. Grarup, I. Strümke, C. Ellervik, M.S. Olesen, et al., Explaining deep neural networks for knowledge discovery in electrocardiogram analysis, Scientific Reports 11 (2021) 1–11.
[20] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
[21] D.L. Kasper, A.S. Fauci, S.L. Hauser, D.L. Longo, J.L. Jameson, J. Loscalzo, Harrison's Principles of Internal Medicine, Vol. 1 & Vol. 2, McGraw Hill Professional, 2018.
[22] Kotikalapudi, R. (2017). keras-resnet. https://github.com/raghakot/keras-resnet. [Online; accessed 9-October-2020].
[23] M. Lerecouvreux, E. Perrier, P. Leduc, O. Manen, M. Monteil, J. Deroche, G. Quiniou, R. Carlioz, Right bundle branch block: electrocardiographic and prognostic features, Archives des maladies du coeur et des vaisseaux 98 (2005) 1232–1238.
[24] T.-Y. Lin, A. RoyChowdhury, S. Maji, Bilinear CNN models for fine-grained visual recognition, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1449–1457.
[25] Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874.
[26] Marinho, L.B., de MM Nascimento, N., Souza, J.W.M., Gurgel, M.V., Rebouças Filho, P.P., & de Albuquerque, V.H.C. (2019). A novel electrocardiogram feature extraction approach for cardiac arrhythmia classification. Future Generation Computer Systems, 97, 564–577.
[27] R.J. Martis, U.R. Acharya, C.M. Lim, J.S. Suri, Characterization of ecg beats from cardiac arrhythmia using discrete cosine transform in pca framework, Knowledge-Based Systems 45 (2013) 76–82.
[28] Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. arXiv preprint arXiv:1706.08947.
[29] B. Pourbabaee, M.J. Roshtkhari, K. Khorasani, Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients, IEEE Transactions on Systems, Man, and Cybernetics: Systems 48 (2018) 2095–2104.
[30] T. Rieg, J. Frick, H. Baumgartl, R. Buettner, Demonstration of the potential of white-box machine learning approaches to gain insights from cardiovascular disease electrocardiograms, PloS one 15 (2020), e0243615.
[31] Sattar, Y., & Chhabra, L. (2020). Electrocardiogram. StatPearls [Internet],.
[32] M. Sharma, R. San Tan, U.R. Acharya, A novel automated diagnostic system for classification of myocardial infarction ECG signals using an optimal biorthogonal filter bank, Computers in biology and medicine 102 (2018) 341–356.
[33] P. Singh, I. Srivastava, A. Singhal, A. Gupta, Baseline wander and power-line interference removal from ECG signals using Fourier decomposition method, in: Machine Intelligence and Signal Analysis, Springer, 2019, pp. 25–36.
[34] Sourajit2110 (2018). Residual-attention-convolutional-neural-network. https://github.com/Sourajit2110/Residual-Attention-Convolutional-Neural-Network. [Online; accessed 7-February-2021].
[35] Strodthoff, N., Wagner, P., Schaeffter, T., & Samek, W. (2020). Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. arXiv preprint arXiv: 2004.13701.
[36] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In: International Conference on Machine Learning (pp. 3319–3328). PMLR.
[37] S.S. Swain, D. Patra, Y.O. Singh, Automated detection of myocardial infarction in ecg using modified stockwell transform and phase distribution pattern from time-frequency analysis, Biocybernetics and Biomedical Engineering 40 (2020) 1174–1189.
[38] Tavares, C. d. A.M., Samesima, N., Hajjar, L.A., Godoy, L.C., Padrão, E.M.H., Neto, F.L., Facin, M., Jacob-Filho, W., Farkouh, M.E., & Pastore, C.A. (2021). Clinical applicability and diagnostic performance of electrocardiographic criteria for left ventricular hypertrophy diagnosis in older adults. Scientific Reports, 11, 1–10.
[39] O.K. Utomo, N. Surantha, S.M. Isa, B. Soewito, Automatic sleep stage classification using weighted ELM and PSO on imbalanced data from single lead ECG, Procedia Computer Science 157 (2019) 321–328.
[40] M. Velasco, E. Rojas, Non–Q-Wave Myocardial Infarction: Comprehensive Analysis of Electrocardiogram, Pathophysiology, and Therapeutics, American Journal of Therapeutics 20 (2013) 432–441.
[41] P. Wagner, N. Strodthoff, R.-D. Bousseljot, D. Kreiseler, F.I. Lunze, W. Samek, T. Schaeffter, PTB-XL, a large publicly available electrocardiography dataset, Scientific data 7 (2020) 1–15.
[42] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3156–3164.
[43] 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.
[44] M. Wasimuddin, K. Elleithy, A. Abuzneid, M. Faezipour, O. Abuzaghleh, Stages-based ecg signal analysis from traditional signal processing to machine learning approaches: A survey, IEEE Access (2020).
[45] P.K. Whelton, R.M. Carey, W.S. Aronow, D.E. Casey, K.J. Collins, C. Dennison Himmelfarb, S.M. DePalma, S. Gidding, K.A. Jamerson, D.W. Jones, et al., 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines, Journal of the American College of Cardiology 71 (2018) e127–e248.
[46] Y.-C. Yeh, W.-J. Wang, C.W. Chiou, A novel fuzzy c-means method for classifying heartbeat cases from ECG signals, Measurement 43 (2010) 1542–1555.
[47] O. Yildirim, M. Talo, E.J. Ciaccio, R. San Tan, U.R. Acharya, Accurate deep neural network model to detect cardiac arrhythmia on more than 10,000 individual subject ecg records, Computer methods and programs in biomedicine 197 (2020), 105740.
[48] J. Zheng, J. Zhang, S. Danioko, H. Yao, H. Guo, C. Rakovski, A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients, Scientific data 7 (2020) 1–8.