International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Heart Murmur Detection with Phonocardiogram Recordings: Analysis of Ensemble Learning Model Performance within XAI Framework

Sandhya Samant[1*], Dr. Amit Dixit[2]

Submitted: 25/12/2023 **Revised**: 05/02/2024 **Accepted**: 15/02/2024

Abstract: Accurate and reliable information on human heart health is key to its prognosis. Most recently, advanced machine learning and deep learning methods are aiding the doctors in decision making. However, it still evades them to understand how a ML or DL model is able to do so. This calls for use of ML/DL model performance interpretation frameworks to correlate a particular model's performance with its internal architecture and functioning. In this study, an attempt is made to interpret the performance two different classification models that participated in the Physionet Challenge 2022. SHAP XAI framework-based interpretations of the performances of one heart murmur detection model and one clinical outcome prediction model are done. The heart murmur detection model selected for interpretation is a transformer-based deep neural network (T-DNN) whereas the clinical outcome prediction model selected for interpretation is a Random Forest boosted with AdaBoost boosting strategy. The dataset considered for model performance interpretations is the *CirCor DigiScope* dataset. The dataset contains phonocardiogram recordings, socio-demographic information and other auxiliary information. The T-DNN is trained on DWT features computed from segmented phonocardiogram signals for three-class (Present, Absent, and Unknown) classification task. The AdaBoost-RF is trained on collection of features including statistical measures, wavelet transform features, time-based and frequency-based features. ANOVA method is used to reduce the dimensionality of the total number of features to 110. The AdaBoost-RF performs a binary (Normal and Abnormal) classification task. The T-DNN model performed classification with overall accuracy of 90.23% whereas the AdaBoost-RF model performed classification with overall accuracy of 89.1%. Shapley importance plot, summary plot and Swarm charts are used to interpret the classification performance of the T-DNN and the AdaBoost-RF here. The study provides insights into the workings of advanced machine learning and deep learning models during detection and identification of heart health from phonocardiogram recordings.

***Keywords**:* Multi-source phonocardiograms, SHAP XAI, Digi Scope dataset, ensemble learning

1. Introduction

Heart disease encapsulates a range of conditions that affect the heart. These conditions can include heart attack, heart failure, coronary artery disease (narrowing of the arteries), arrhythmias (irregular heartbeat), and others, These conditions can adversly affect the heart's structure and function. Heart defects or abnormalities that are present at birth are termed as Congenital heart disease (CHD). These defects can affect the heart's walls, valves, or blood vessels, disrupting normal blood flow through the heart. Rheumatic heart disease (RHD) is a condition that develops as a complication of untreated or inadequately treated streptococcal throat infection, specifically caused by group A streptococcus bacterium. This infection can lead to rheumatic fever, which in turn can cause inflammation and damage to the heart valves and other heart structures. Both CHD and RHD diagnosis typically involves a combination of medical history, physical examination, blood tests, electrocardiogram (ECG), echocardiogram, stress tests, and cardiac catheterization but the cardiac health pre-screening is almost alwats done by observing cardiac auscultation via a stethoscopes. Observation and interpretation of heartbeat sounds is an '*Art*' as it is '*Science*'. Therefore, there has been critical discussions on what is the best prcatice to skill the task. Most importanly, it needs acute hearing state which is a human realted parameter and cannot be controlled with precision. With the significant advances in instrumentation technology, digital phonocardiography is now a powerful assistive tool in heart health monitoring. Phonocardiography is used to study the various sounds produced by the heart during its cycle of contraction and relaxation. A phonocardiogram (PCG) can be acquired via a combination of (a).High-fidelity stethoscope front-ends and (b). High-resolution digital sampling circuitry. The stethoscope front-ends act as a diaphragm or a membrane that vibrates when it comes into contact with heartbeat sound waves. These vibrations represent the acoustic pressure waves which are registered as a discrete-time signal by the sampling circuitry. These discrete-time signals can further be interpreted by the use of detection algorithms. A phonocardiogram typically consists of several key components:

- S1 and S2 sounds: S1 represents the sound of the closure of the mitral valve (MV) and the tricuspid valve (TV)

[1*]*Ph.D Scholar, Quantum University,Roorkee*
[2]*Dean Ph.D Programme, Quantum University,Roorkee*

during systole, and S2 represents the sound of theclosure of the aortic valve (AV) and the pulmonary valve (PV) during diastole. The positions of these valves in 2D is shown in Figure 1.

- Murmurs: The turbulent blood flow through the heart or blood vessels, cause abnormal sounds. These abnormal sounds are termed as 'Murmurs' that include valve abnormalities or other structural issues.

- Other Sounds: Phonocardiograms may also capture other sounds such as clicks (e.g., in mitral valve prolapse), snaps (e.g., in aortic stenosis), and rubs (e.g., pericardial friction rub).

Compared to auscultation (listening to heart sounds with a stethoscope), phonocardiography provides a more detailed and objective assessment of heart sounds. It allows for precise measurement and analysis of sound characteristics, which can aid in diagnosing subtle abnormalities that may not be easily detected by auscultation alone.

Feature extraction and selection are common processes witin machine learning approaches used for murmur detection. Mel-frequency cepstral coefficients (MFCC) are most commonly used in speech and audio processing and can be adapted for heart sound analysis [1][2]. MFCCs are robust against noise and other distortions which making them suitable for real-world applications. They capture the frequency characteristics of heart sounds, which are crucial for distinguishing murmurs from normal heart sounds. By reducing the dimensionality of the feature space, MFCCs simplify the task of classification without losing significant information [3]. Their integration with machine learning algorithms contributes to advancing automated and accurate murmur detection systems in clinical settings [4][5]. Wavelet Transform is another feature extraction and selection technique that is useful for decomposing heart sound signals into different frequency components, which can then be analyzed to identify specific patterns associated with murmurs [6]. This decomposes a signal into different frequency components with varying resolutions in time. Unlike the Fourier transform, which provides a fixed resolution in both time and frequency, wavelet transform adapts to local characteristics of the signal [7]. In the context of murmur detection from heart sound recordings, the recordings can be decomposed using wavelet transform into wavelet coefficients at different scales. This decomposition allows capturing both low-frequency components (e.g., heartbeats) and high-frequency components (e.g., murmurs) separately [8]. Machine learning algorithms can then be applied to these wavelet coefficients to classify heart sounds as normal or abnormal (murmurs)[8]. The patterns in the wavelet coefficients associated with murmurs (such as specific frequency distributions or transient spikes) can be learned by the models [9]. It has been also repored that rather than the use of MFCC or wavelet transform, statistical features such as mean, standard deviation, skewness, kurtosis, and spectral entropy can provide valuable information about the characteristics of heart sounds, aiding in the detection of abnormalities [10]. However, in this study, wavelet transfoorm features, statistical measures, time-nased and frequency-based features are computed and used as features.

Artificial intelligence or AI has been increasingly utilized to detect cardiac murmurs due to its ability to analyse large amounts of data quickly and accurately. Advanced machine learning models are increasingly being developed and utilized for the detection of cardiac murmurs. These models leverage various techniques and datasets to accurately identify abnormalities in heart sounds. Deep learning models based on findational architectures such as the recurrent neural networks (RNNs) and the convolutional neural networks (CNNs), can be trained on large datasets of labeled heart sound recordings[11][12] to perform heart disease detection or identification. CNNs are particularly effective for extracting useful information directly from waveforms of heart sounds or from features such as wavelet transform features or they can directly process spectrograms of heart sound recordings [13]. Given the limited availability of labeled heart sound datasets, data augmentation techniques can help increase the diversity of training samples and improve the model's robustness [14]. The RNNs can capture temporal dependencies in the sound sequences [15]. However, if the sequence is too long, it makes learning in the RNNs a computationally expensive and cumbersome. Therefore state-of-art techniques that use combination of the CNNs and the RNNs have been proven to learn spatial, spectral, and time-dependent characteristics of heart murmur from PCG waveform simultaneously [16]. Also, transfer learning has been explored for heart murmur detection. In transfer learning, CNN models, pre-trained on large audio datasets, has been fine-tuned for murmur detection [17]. This strategy approach leverages learned features from general audio patterns, potentially enhancing performance with minimal data. However, learning the patterns of a heartbeat (normal or abnormal) can be challenging among the range of patterns present in the audio signals. Most recently, transformer-based DNNs have been efficient and effective in numeruos time-series signal processing applications such as the PCG signals at hand here. Transformers have been effective DNN architectures in natural language processing (NLP) applications[18]. These are an advanced version of the conventional CNNs and RNNs because of its ability to capture global dependencies in the input through self-attention and cross-attention

mechanisms [18]. These mechanisms force the DNN to learn from regions of most significant impact on the class prediction. Transformer-based DNNs are gaining popularity in medical applications[18].

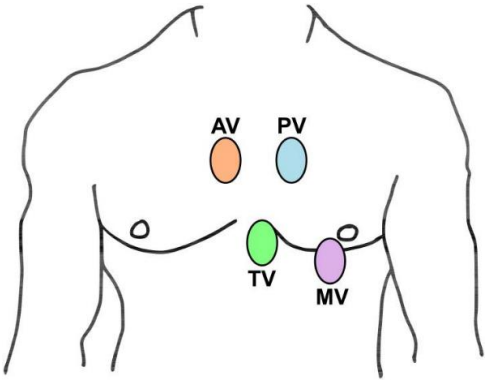


Fig 1 *Positions of the heart valves in 2D.*

Alternatively to deep learning techniques, advanced machine learning algorithms such as the XGBoost, LightGBM, RUSBoost are also powerful for ensemble learning, combining the predictions of multiple weak learners such as the decision trees to improve accuracy in classification tasks like murmur detection [2][19][20]. Furthermore, several versions of boosting were tested on the PCG data. Models are typically evaluated using metrics such as accuracy, sensitivity, specificity, and area under the ROC curve (AUC), reflecting their ability to correctly classify heart sound recordings. However, model performance evaluations of DL models with these idicators is not enough since these models are a blackbox and it is difficult to pinpoint how much a particular component is contributing towarsd its performance. Therefore explainable AI or XAI frameworks are gaining attention in DL models' performance interpretations. Therefore, to understand the impact of model architectures on heart murmur detection and classification performance, this study explores the potential of a transformer-based DNN and a boosted RF classifier within an XAI framework. The transformer-based DNN selected for analysis is proposed in [21] and the boosted RF selected for analysis is proposed in [22]. The Shapley XAI framework analyses the model performance of both the classifiers.

In this study, an attempt is made to interpret the performance two different classification models that participated in the Physionet Challenge 2022 [23, 24]. SHAP XAI framework-based interpretations of the performances of one heart murmur detection model and one clinical outcome prediction model are done. The heart murmur detection model selected for interpretation is a transformer-based deep neural network (T-DNN) proposed in [21] whereas the clinical outcome prediction model selected for interpretation is a Random Forest boosted with AdaBoost boosting strategy proposed in [22]. The dataset considered for model performance interpretations is the *CirCor DigiScope* dataset [25]. The dataset contains phonocardiogram recordings, socio-demographic information and other auxiliary information. The T-DNN is trained on DWT features computed from segmented phonocardiogram signals for three-class (Present, Absent, and Unknown) classification task. The AdaBoost-RF is trained on collection of features including statistical measures, wavelet transform features, time-based and frequency-based features. ANOVA method is used to reduce the dimensionality of the total number of features. The AdaBoost-RF performs a binary (Normal and Abnormal) classification task. Shapley importance plot, summary plot and Swarm charts are used to interpret the classification performance of the T-DNN and the AdaBoost-RF here. The rest of the paper is divided into sections. Section 2 and its sub-sections incorporates the materials and methods utilized in the study. Section 3 provides results obtained and its discussion. Section 4 concludes the study.

2. Materials and Methods

2.1 Dataset: Description and Preparation: CirCor DigiScope dataset

The dataset for this paper is the CirCor DigiScope dataset which consolidated via a series of campaigns held in various locations [25]. The dataset was collected as part of two mass screening campaigns conducted in Northeast Brazil in July-August 2014 and June-July 2015. The data collection was approved by the 5192-Complexo Hospitalar HUOC/PROCAPE institutional review board, under the request of the Real Hospital Portugues de Beneficencia em Pernambuco. The campaign was termed "Caravana do Coração" (Portuguese for "Caravan of the Heart"). In this campaign, a total of 2061 participants participated. From the original 2061 participants, 493 participants were excluded for not meeting the eligibility criteria. The remaining 1568 participants underwent a clinical examination (anamnesis and physical examination). A nursing assessment (physiological measurements), and cardiac investigations (chest

radiography, electrocardiogram, and echocardiogram) were made during the examination. The participants also completed a socio-demographic questionnaire that was later used to annotate the recordings. During these campaigns a total number of 5272 heart sound recordings were collected. High-fidelity stethoscope front-ends are placed at four auscultation locations (PV, AV, TV, and MV) of 1568 subjects. The subjects were aged between 0 and 21 years. The mean age in the dataset was 6.1 years and the standard deviation was 4.3 years. The duration of observation ranged between 4.8 to 80.4 seconds. The mean observation time was 22.9 seconds and the standard deviation was 7.4 seconds. The overall recording time was approximately 33.5 hours. A human annotator labelled each cardiac murmur in the dataset. The annotations include time lapsed during the murmur event, shape of the PCG during the event, location, pitch, grading, and quality of the recording. The dataset majorly came from six sources listed as a, b, c, d, e, and f here (refer Table 1.

Imbalanced data

Another major issue with multi-source database is the imbalance in sample proportions which can adversely affect the performance of any machine learning oriented algorithm [26]. Here, a basic technique of repeating the minority class samples to have a balanced class proportion i.e.SMOTE is used. Table _ lists the number of samples collected for each class.

Table 1 *Number of samples collected from different data source to balance the class sample proportion.*

Source	Number of Normal records	Number of Pathological records
a	1840	1564
b	1782	1848
c	1755	1752
d	1848	1584
e	1930	1920
f	1871	1830
Total	11,026	10,498

2.2 Heart Murmur Detection: Transformer-Based DNN

Input data: The model used four-channel PCG data from 942 patients for training. The order of channels is AV, MV, PV, and TV. The first 40 seconds of the recordings is used and 'same' padding is used wherever necessary.

Feature extraction: The heart function characteristics are embedded within the recorded PCG signals along with other contaminated signals. a discrete wavelet transform (DWT) was employed on the original PCG signals. Each PCG channel signal comprised of 160,000 samples which is equivalent to 40 seconds length, was fragmented into 5,000 segments. Each segment is therefore consisting 32 samples per channel. Further, the 32-sample segment of each PCG channel is represented by a total of 30 features. The features from each channel are concatenated to form a wavelet power transformation. In the end, a transformed signal of 20,000 samples (4 channels × 5000 32-sample segments) is achieved. This transformed signal is used to train the DNN.

Transformer-based DNN: Their transformer-based DNN proposed in [21] is based on four major components. The first component is a feature encoder which is based on a set of two one-dimensional convolutions acting on transformed input data. These one-dimensional convolutions act as an encoder that produces features of lower dimensional space. This encoder utilizes Gaussian error linear unit (GELU) as an activation function. Batch normalization and pooling is also used. The encoder outputs a feature space of size 30×1250. Next, the second component is a positional encoder that encodes the feature space from the feature encoder via a series of *sin* and *cosine* representations. The third components is a transformer that uses multi-head attention mechanism. This transformer performs a scaled dot product on the sin-cosine representations from the positional encoder. The DNN uses a single transformer unit and provides a single attention vector. Finally, the forth component is a decoder consisting of a series of fully connected and pooling layers that decodes the attention vector from the transformer layer and produces a single

value per input. Figure 2 presents the layout of the four-component DNN. More details of this DNN architecture are discussed in [21].

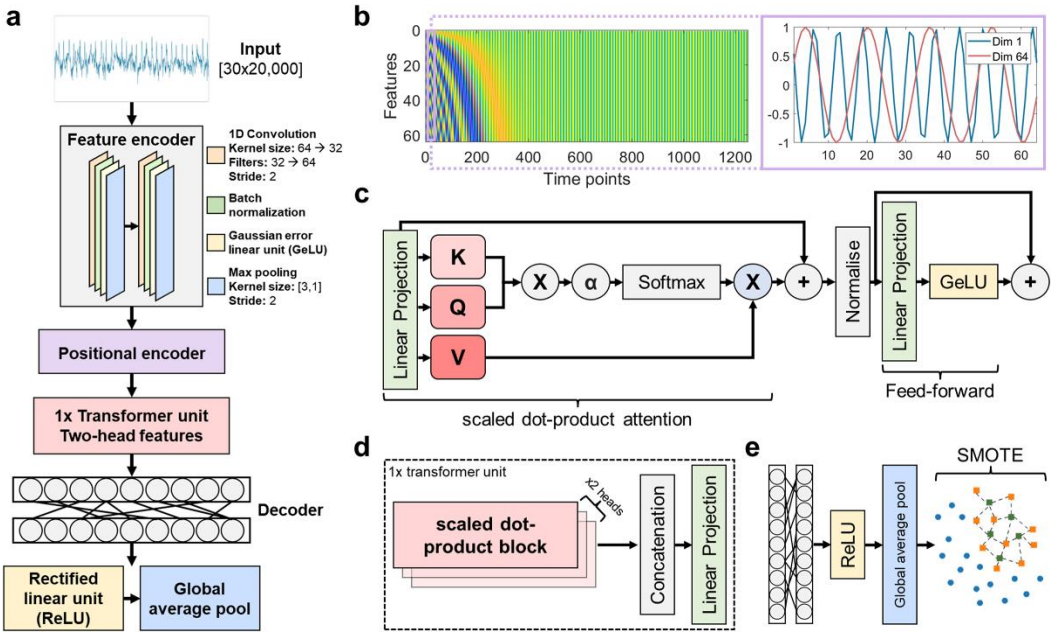


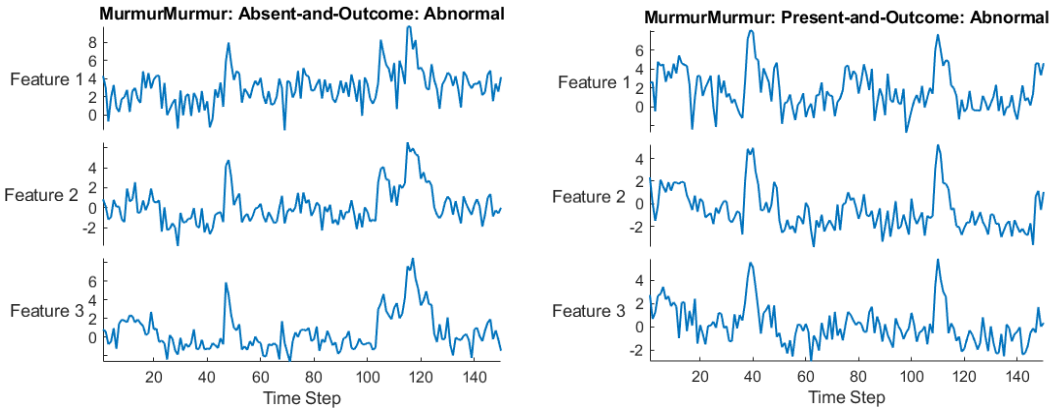
Fig 2 *Figure showing the hybrid architecture of the four-component DNN (T-DNN) proposed in [21].*

2.3 Clinical Outcome: Ensemble learning: Boosted Random Forest

The classier selected here for evaluation for the clinical outcome classification task is the Random Forest (RF) classifier boosted with AdaBoost (Adaptive Boosting) technique as proposed in [22].

Features: Each recording was processed the same way, after preprocessing and segmentation S1, systole, S2, and diastole sections were separated and the same features were calculated for each heart cycle segment along with certain features for the entire signal. From each of these channel signals, 900 wavelet-based features and 780 statistical-based, time-based, and frequency-based features are computed.

Feature selection: Analysis of variance ANOVA based feature selection was used to select the 110 best scoring features from the 900 (wavelet) + 780 (statistical, time, and frequency) features. Figure 3 shows sample features obtained from a sample PCG. A few prominent features computed are listed in table 2.



(a) **(b)**

Fig 3 *Transformed features (1, 2, 3) for first 150 samples and class, (a) Murmur-Absent, Outcome-Abnormal (b) Murmur-Present, Outcome-Abnormal.*

Table 2 List of top few features selected via the ANOVA method.

Prominent features	Prominent features
Mean	Median of the Histogram Counts of the Beat
Average Magnitude of the Fourier Transform	Wavelet Entropy
Minimum of the Fall-time	High value of the Beat State-level
Average Peak value of the beat	Median of the Pulse-Width
Minimum of the Pulse-Period	Median of the Fall-time
Third Order Moment	Median of the Pulse-Period
Minimum of the Pulse-Width	Average Angle of the Fourier Transform
Median of the beat overshoot	Kurtosis
Low value of the Beat State-level	Standard Deviation

Classifier: Random Forest with AdaBoost

The classier selected here for evaluation for the clinical outcome classification task is the Random Forest (RF) classifier boosted with AdaBoost (Adaptive Boosting) technique as proposed in [22]. RF is an ensemble learning technique composed of an ensemble of smaller decision trees. Each tree performs classification on a bootstrapped random subset of the data, using a random subset of the features. AdaBoost focuses on correcting the mistake of decision trees (weak classifiers) by iteratively reweighting the training samples. The technique works as follows.

- Assign weights to each sample based on $w_i = \frac{1}{N}$, where N is the number of samples.
- At a particular iteration t:

1. Train a decision tree $h_t(x)$ on the weighted dataset.

2. Compute the error for samples that are incorrectly classified as follows.

$$\varepsilon_t = \frac{\sum_{i=1}^{N} w_i . 1[h_t(x_i) \neq y_i]}{\sum_{i=1}^{N} w_i} \quad (1)$$

3. Compute the weight for the decision tree: $\alpha_t = \frac{1}{2} ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$

4. Update the sample weight: $w_i \leftarrow w_i . exp\left(-\alpha_t y_i h_t(x_i)\right)$

5. Normalize $w_i$so that $\sum w_i = 1$

- Final prediction: $H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$

Boosting Integration: Sequentially, errors from previous trees are minimized, like in boosting. Mathematically, the predications in Boosted RF are updated as:

$$F_t(x) = F_{t-1}(x) + \eta \sum_{b=1}^{B} \alpha_b . f_b(x) \quad (2)$$

Where

- $f_b(x)$is the b^{th} tree from the random forest.
- α_b is the weight assigned to each tree.
- η is the learning rate.

The loss function used for learning in boosted RF is a log-loss which is governed by;

$$L\big(y, F(x)\big) = -\sum_{i=1}^{N} \big[y_i \, log\, F\,(x_i) + (1 - y_i)\, log\big(1 - F(x_i)\big)\big] \quad (3)$$

The boosted RF had a learning rate of 0.1 and 60 learners in its ensemble, with maximum branching set to 20. In clinical outcome classification, misclassified ”Abnormal” cases had a cost of 2.

3 Experiment Setup, Results, and Discussion

3.1 Murmur detection: Transformer-based DNN

The transformer-based DNN proposed in [21] used the hyperparameter settings according to Table 3. The model uses Adam as optimizer and an initial learning rate of 0.001. The model is trained for 60 epochs and the learning rate decays 10% after the 40th epoch. The model uses a 10-fold cross-validation strategy for more optimal and generalized model performance. Details of hyperparameter settings for T-DNN model is listed in Table 3. Sample distribution percentages for training, validation, and testing sets is listed in Table 4. During

model training, the model employs synthetic minority oversampling technique (SMOTE) to balance out the imbalances in class proportions in the training set. Class proportions before and after SMOTE is applied is listed in Table 5. The trained T-DNN model when tested on testing set provided an overall accuracy of 90.23% while the specificity and sensitivity achieved are 70.22% and 70.41% respectively as listed in Table 6.

Table 3 Hyperparameter settings for the transformer-based DNN (T-DNN).

Hyperparameter	Setting
Model name	T-DNN
Optimizer	Adam
Epochs	60
Learning rate	0.001
Learning rate decay	10% at the 40^{th} epoch
Cross-validation	10-fold
Minority class over sampling technique	SMOTE

Table 4 *Distribution of dataset samples across training, validation, and testing sets.*

Sample distribution percentage during T-DNN model training		
Training	**Validation**	**Testing**
0.8	0.1	0.1

Table 5 *Class proportion distribution correction with SMOTE.*

Class proportions in the training dataset (before SMOTE)		
Present	**Absent**	**Unknown**
0.74	0.19	0.07
Class proportions in the training dataset (after SMOTE)		
Present	**Absent**	**Unknown**
0.44	0.29	0.27

Table 6 *T-DNN model performance metrics*

Model	**Performance scores (testing set) in %**		
	Accuracy	**Specificity**	**Sensitivity**
T-DNN	90.23	70.22	72.41

To interpret model performance, SHAP XAI framework is employed. Figure 4 presents a plot between predictors and its corresponding Shapley value. It is evident from Figure 4 that predictors id 80-100 are contributing significantly for all three classes. More specifically, predictor id 85 is significantly contributing towards identification of the *Unknown* class, predictor id 105 is significantly contributing towards identification of the *Present* class and, predictor id 91 is significantly contributing towards identification of the *Absent* class. Figure 5 shows mean absolute Shapley values for top 10 predictors in descending order. For example, the topmost predictor has the highest value (for *Present* class). However, this predictor contributes less towards the *Absent* class than predictor at position 2. For the *Unknown* class, predictor at position 3 is contributing the most but its contribution value is less than what predictor at position 1 is contributing towards the *Present* class. This

is why it is positioned at 3. The position of the predictors in the order of contribution towards the *Present* class is- 1, 2, 7, 8, 4, 3, 5, 6, 10, and 9. Whereas the position of the predictors in the order of contribution towards the *Absent* class is- 2, 3, 4, 7, 8, 10, 1, 6, 5, and 9 and the position of the predictors in the order of contribution towards the *Unknown* class is- 1, 5, 6, 2, 9, 3, 4, 7, 8, and 10. Please note these are the ranking for top 10 predictors among the 110 predictors.

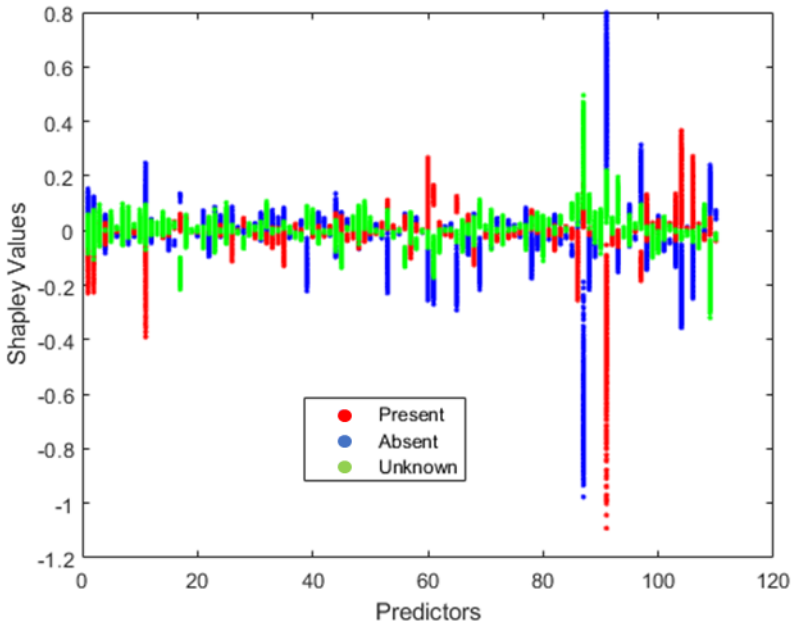


Fig 4 *Shapley values for each predictor; red- Present class, blue- Absent class, and green- Unknown class.*

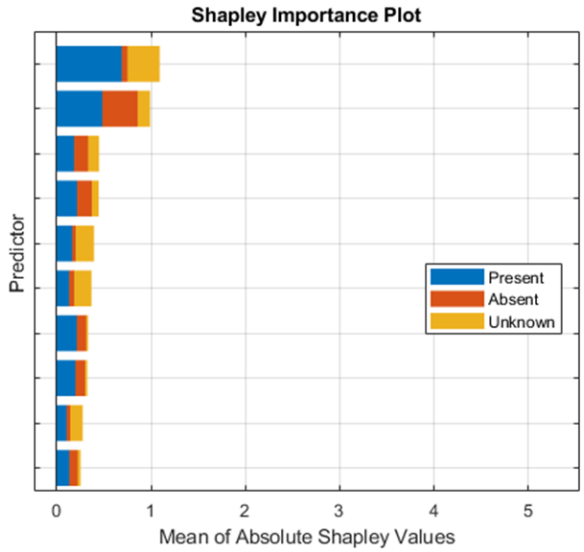


Fig 5 *Mean absolute Shapley values for top 10 predictors; blue- Present class, red- Absent class, and orange- Unknown class.*

Figure 6 shows the Swarm chart for the *Present* class. In this figure, higher values of predictors are indicated in red color and lower values of predictors are indicated by blue colour. Predictor at position 1 and 2 show high absolute Shapley values whereas predictor 8, 9, and 10 show low absolute Shapley values. Figure 6 also shows sample distributions contributing towards the *Present* class. For example, the predictor at position 1 (topmost) has absolute Shapley values for most samples is in the range 0.7- 0.9 or the predictor at position 2 (second from top) has absolute Shapley values for most samples is in the range 0.3- 0.5.

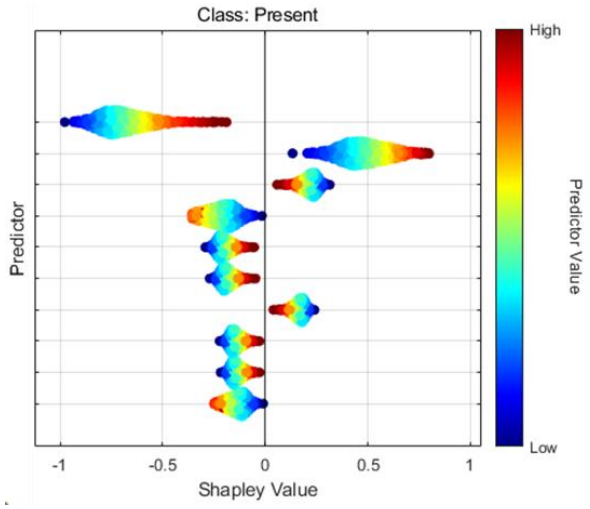


Fig 6 *Swarm chart of top 10 predictors for the Present class.*

Figure 7 shows the Swarm chart for the *Absent* class. In this figure, higher values of predictors are indicated in red color and lower values of predictors are indicated by blue colour. Predictor at position 1 and 2 show high absolute Shapley values whereas predictor 8, 9, and 10 show low absolute Shapley values. Figure 7 also shows sample distributions contributing towards the *Present* class. For example, the predictor at position 1 (topmost) has absolute Shapley values for most samples is in the range 0.2- 0.6 or the predictor at position 2 (second from top) has absolute Shapley values for most samples are around 0.2.

Figure 8 shows the Swarm chart for the *Unknown* class. In this figure, higher values of predictors are indicated in red color and lower values of predictors are indicated by blue colour. Predictor at position 1 and 2 show high absolute Shapley values whereas predictor 8, 9, and 10 show low absolute Shapley values. Figure 8 also shows sample distributions contributing towards the *Present* class. For example, the predictor at position 1 (topmost) has absolute Shapley values for most samples is in the range 0.3 - 0.5 or the predictor at position 2 (second from top) has absolute Shapley values for most samples are around 0.3 – 0.1.

Interpreting Figure 6, 7, and 8 simultaneously indicates that for the *Present* class, positive Shapley values are obtained with higher predictor values for predictor at position 2, 3, and 7. Whereas negative Shapley values are obtained with higher predictor values for predictor at position 1, 4, and 10. For the *Absent* class, positive Shapley values are obtained with higher predictor values for predictor at position 5, 6, and 8. Whereas negative Shapley values are obtained with higher predictor values for predictor at position 1 and 7.

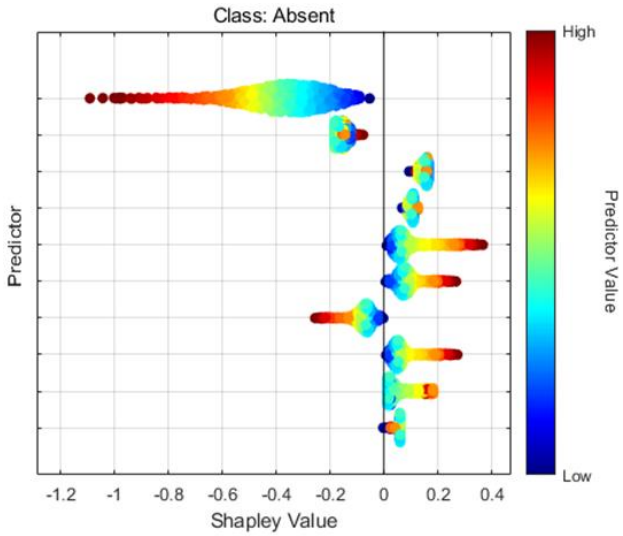


Fig 7 *Swarm chart of top 10 predictors for the Present class.*

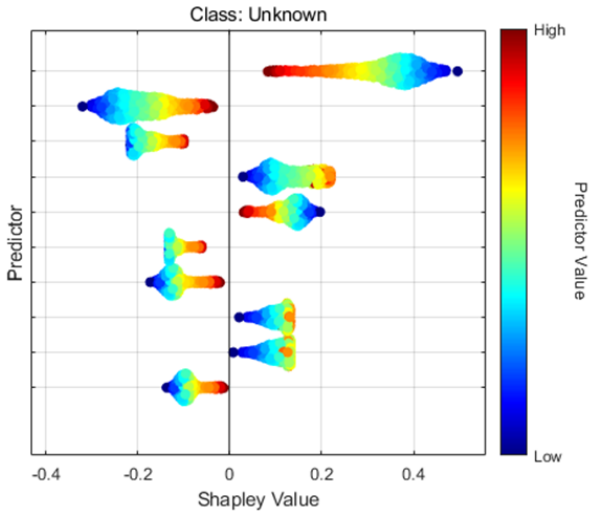


Fig 8 *Swarm chart of top 10 predictors for the Present class.*Clinical outcome: AdaBoost-RF

The AdaBoost-RF proposed in [22] used the hyperparameter settings according to Table 7. The model uses 60 weak learners, a learning rate of 0.1, and a maximum branching set of 20. A voting mechanism is used to aggregate the classifications from different weak learners (decision trees). Table 8 lists the sample proportion distribution for the training, validation, and testing set. Table 9 lists the class proportions for the *Normal* and the *Abnormal* clinical outcome class. It is evident that the class proportions for the *Normal* and the *Abnormal* class are equivalent therefore SMOTE is not required. The trained AdaBoost-RF model when tested on the testing set provided an overall accuracy of 89.1% while the specificity and sensitivity achieved are 86.6% and 91.6% respectively as listed in Table 10.

Table 7 *Hyperparameter settings for AdaBoost-RF.*

Hyperparameter	**Setting**
Learning rate	0.1
Number pf weak learners	60
Maximum branching set	20

Table 8 *Distribution of dataset samples across training, validation, and testing sets.*

Sample distribution percentage during RF-AdaBoost training		
Training	**Validation**	**Testing**
0.8	0.1	0.1

Table 9 *Class proportion distribution.*

Class proportions in the training set	
Normal	**Abnormal**
0.48	0.52

Table 10 *AdaBoost-RF model performance on testing set.*

Model	**Performance scores (testing set) in %**		
	Accuracy	**Specificity**	**Sensitivity**
AdaBoost-RF	89.1	86.6	91.6

To interpret model performance, SHAP XAI framework is employed. Figure 9 presents a plot between predictors and its corresponding Shapley value. It is evident from Figure 9 that predictors id 50-80 are contributing most significantly for both the classes. More specifically, predictor id- *62* is most significant the towards identification of the *Noraml* and the *Abnormal* class and predictors id- *50*, *62*, and *80* are significantly contributing towards identification of the *Abnormal* class. Figure 10 shows mean absolute Shapley values for top 10 predictors in descending order. For example, the topmost predictor has the highest value for the *Normal and* the *Abnormal* class. The position of the predictors in the order of contribution towards the *Normal* class is- 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Whereas the position of the predictors in the order of contribution towards the *Abnormal* class is- 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. Please note these are the ranking for top 10 predictors among the 110 predictors.

Figure 11 shows the Swarm chart for the *Normal* class. In this figure, higher values of predictors are indicated in red color and lower values of predictors are indicated by blue colour. Predictor at position 1 and 6 show high absolute Shapley values whereas predictor 7, 8, and 9 show low absolute Shapley values. Figure 12 also shows Swarm chart for the *Abnormal* class. For example, the predictor at position 4 (from top) provides negative Shapley values for higher values of the predictor and provides positive Shapley values for lower values of the predictor. Interpreting Figure 11 and 12 simultaneously indicates that for the *Normal* class, positive Shapley values are obtained with higher predictor values for predictor at position 4, 6, and 10. Whereas negative Shapley values are obtained with higher predictor values for predictor at position 1, 6, and 9.

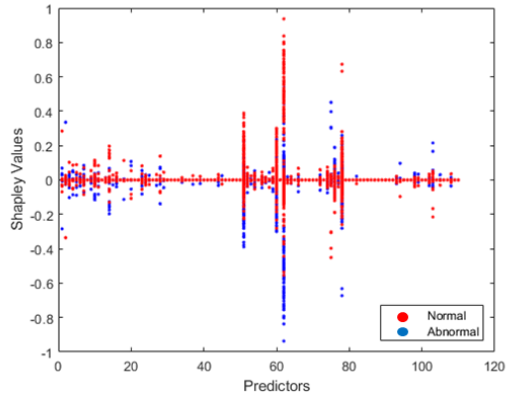


Fig 9 *Figure showing Shapley values corresponding to each predictor.*

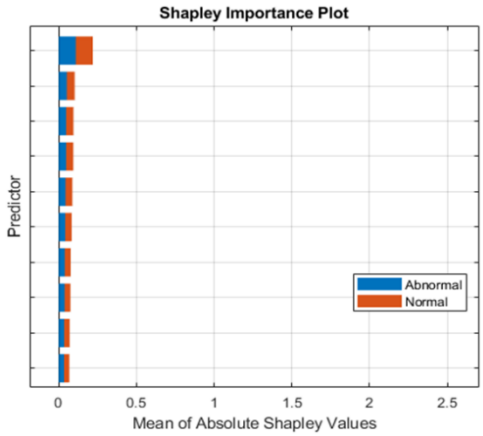


Fig 10 *Mean absolute Shapley values for top 10 predictors; blue- Abnormal class and, red- Normal class.*

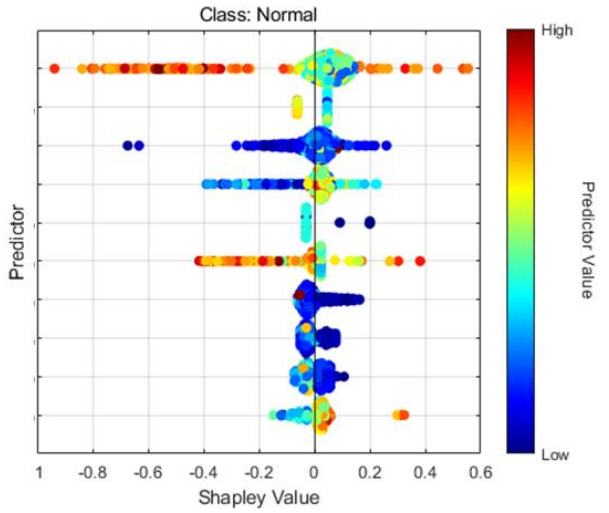


Fig 11 *Swarm chart of top 10 predictors for the clinical outcome −Normal class.*

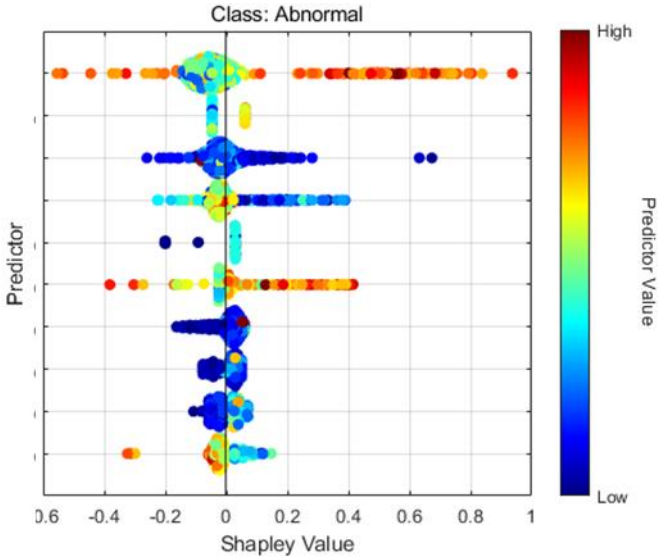


Fig 12 *Swarm chart of top 10 predictors for the clinical outcome −Abnormal class.*

4 Conclusion

In this study, SHAP XAI framework-based interpretations of the performances of one heart murmur detection model and one clinical outcome prediction model are done. The heart murmur detection model selected for interpretation is a transformer-based deep neural network (T-DNN) whereas the clinical outcome prediction model selected for interpretation is a random forest boosted with AdaBoost. The dataset considered for model performance interpretations is the *CirCor DigiScope* dataset. The dataset contains phonocardiogram recordings, socio-demographic information and other auxiliary information. The T-DNN is trained on DWT features computed from segmented phonocardiogram signals for three-class classification task. The three classes are murmur−*Present*, murmur−*Absent*, and murmur−*Unknown*. The imbalance in class-wise training sample proportions is balanced via SMOTE method. The T-DNN model performed classification with overall accuracy of 90.23%. The Shapley values obtained from the SJAP XAI framework-based interpretation of the T-DNN reflects which features are critical or are contributing significantly towards a particular class. Shapley importance plot, summary plot and Swarm charts are used to interpret the classification performance of the T-DNN here. The AdaBoost-RF is trained on collection of features including statistical measures, wavelet transform features, time-based and frequency-based features. ANOVA method is used to reduce the dimensionality of the total number of features to 110. The AdaBoost-RF performs a binary classification task. The two classes are clinical outcome *Normal*, murmur−*Absent*, and clinical outcome *Abnormal*. SMOTE is not required in this scenario since both classes has equivalent sample size. The AdaBoost-RF model performed classification with overall accuracy of 89.1% whereas the specificity and sensitivity of this model are 86.6% and 91.6% respectively. The Shapley values obtained from the SJAP XAI framework-based interpretation of the AdaBoost-RF reflects which features are critical or are contributing significantly towards a particular class. Shapley importance plot, summary plot

and Swarm charts are used to interpret the classification performance of the AdaBoost-RF here. This study provides insights into the workings of advanced machine learning and deep learning models during detection and identification of heart health from phonocardiogram recordings. Similar studies in future could help establish the significance of say DNN architecture or nature of ensemble strategy in heart disease detection tasks.

References

[1] Venkataramani VV, Garg A, Priyakumar UD (2022) Modified Variable Kernel Length ResNets for Heart Murmur Detection and Clinical Outcome Prediction Using Phonocardiogram Recordings. Computing in Cardiology 2022-Septe:1–4. https://doi.org/10.22489/CinC.2022.315

[2] Imran Z,Grooby E,Malgi VV,et al(2022) A Fusion of Handcrafted Feature –Based and Deep Learning Classifiers for Heart Murmur Detection .Computing in Cardiology 2022-septe:1-4.https://doi.org/10.22489/cinC.2022.310

[3] Abdul ZK,AI-Talabani AK(2022) Mel Frequency Cepstral Coeddiciemt and its Applications:AReview.IEEE Access 10:122136-122158.https://doi.org/10.1109/ACCESS.2022.3223444

[4] Chang Y, Liu L, Antonescu C (2022) Multi-Task Prediction of Murmur and Outcome from Heart Sound Recordings. Computing in Cardiology 2022-Septe:7-10. https://doi.org/10.22489/CinC.2022.309

[5] Bai Z, Yan B, Chen X, et al (2022) Murmur Detection and Clinical Outcome Classification Using a VGG-like Network and Combined Time-Frequency Representations of PCG Signals. Computing in Cardiology 2022-Septe: 1-4. https://doi.org/10.22489/CinC.2022.318

[6] Maller K, Goda MA (2022) Heart Murmur Detection in Phonocardiographic Signals Using Breathing Noise Suppression. Computing in Cardiology 2022-Septe:2-5. https://doi.org/10.22489/CinC.2022.280

[7] Nivitha Varghees V, Ramachandran KI (2015) Heart murmur detection and classification using wavelet transform and Hilbert phase envelope. 2015 21st National Conference on Communications, NCC 2015 1-6. https://doi.org/10.1109/NCC.2015.7084904

[8] Petrolis R, Paukstaitiene R, Rudokaite G, et al (2022) Convolutional Neural Network Approach for Heart Murmur Sound Detection in Auscultation Signals Using Wavelet Transform Based Features. Computing in Cardiology 2022-Septe:2-5.https://doi.org/10.22489/CinC.2022.043

[9] Comely AK, Mirsky GM (2022) Heart Murmur Detection Using Wavelet Time Scattering and Support Vector Machines. Computing in Cardiology 2022-Septe:1-4. https://doi.org/10.22489/CinC.2022.251

[10] Touahria R, Hacine-Gharbi A, Ravier P (2023) Feature selection algorithms highlight the importance of the systolic segment for normal/murmur PCG beat classification. Biomedical Signal Processing and Control 86:105288. https://doi.org/10.1016/j.bspc.2023.105288

[11] Xu Y, Bao X, Lam HK, Kamavuako EN (2022) Hierarchical Multi-Scale Convolutional Network for Murmurs Detection on PCG Signals.Computing in Cardiology 2022-Septe: 1-4. https://doi.org/10.22489/CinC.2022.439

[12] Warrick PA, Afilalo J (2022) Phonocardiographic Murmur Detection by Scattering-Recurrent Networks. Computing in Cardiology 2022-Septe:10-13. https://doi.org/10.22489/CinC.2022.408

[13] LiX, Ng GA, Schlindwein FS (2022) Transfer Leaming in Heart Sound Classification using Mel Spectrogram. Computing in Cardiology 2022-Septe: 1-4. https://doi.org/10.22489/CinC.2022.046

[14] Lu H, Yip JB, Steigleder T, et al (2022) A Lightweight Robust Approach for Automatic Heart Murmurs and Clinical Outcomes Classification from Phonocardiogram Recordings. Computing in Cardiology 2022-Septe:4-7. https://do1.org/10.22489/CinC.2022.165

[15] Shin JM, Park SY, Kim HS, et al (2022) Leaming Time-Frequency Representations of Phonocardiogram for Murmur Detection. Computing in Cardiology 2022-Septe: 1-4. https://doi.org/10.22489/CinC.2022.126

[16] Alam S, Banerjee R, Bandyopadhyay S (2018) Murmur Detection Using Parallel Recurrent & Convolutional Neural Networks

[17] Costa JL, Couto P, Rodrigues R (2022) Multitask and Transfer Learning for Cardiac Abnormality Detections in Heart Sounds. Computing in Cardiology 2022-Septe: 1-4. https://doi.org/10.22489/CinC.2022.193

[18] Jain SM (2022) Introduction to transformers for NLP. Springer

[19] Walker B, Krones F, Kiskin I, et al (2022) Dual Bayesian ResNet: A Deep Leaming Approach to Heart Murmur Detection. Computing in Cardiology 2022-Septe: 1-4. https://doi.org/10.22489/CinC.2022.335

[20] Summerton S,Wood D,Murphy D,et al(2022) Two Stage Classification for Detecting Murmurs from Phonocardiograms Using Deep and Expert Features .Computing in Cardiology 2022-septe 3-6 https://doi.org/10.22489/CinC.2022.322

[21] Alkhodari M, Azman SK, Hadjileontiadis LJ, Khandoker AH (2022) Ensemble Transformer-Based Neural Networks Detect Heart Murmur in

Phonocardiogram Recordings. In: Computing in Cardiology. IEEE Computer Society

[22] Baydoun M, Safatly L, Ghaziri H, El Hajj A (2020) Analysis of heart sound anomalies using ensemble leaming. Biomed Signal Process Control 62:. https://doi.org/10.1016/j.bspc.2020.102019

[23] Goldberger AL, Amaral LAN, Glass L, et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101:e215-e220

[24] Reyna MA, Kiarashi Y, Elola A, et al (2023) Heart murmur detection from phonocardiogram recordings: The george b. moody physionet challenge 2022. PLOS Digital Health 2:e0000324

[25] Oliveira J, Renna F, Costa PD, et al (2021) The CirCor DigiScope dataset: from murmur detection to murmur classification. IBEE J Biomed Health Inform 26:2524-2535

[26] Deng F, Tu S, Xu L (2021) Multi-source unsupervised domain adaptation for ECG classification. In: 2021 IEEE International Conference on Bioinformatics and Biomedicine(BIBM).pp854-859