MD-CardioNet: A Multi-Dimensional Deep Neural Network for Cardiovascular Disease Diagnosis From Electrocardiogram

Article *in* IEEE Journal of Biomedical and Health Informatics · August 2023

DOI: 10.1109/JBHI.2023.3308856

CITATIONS
2

READS
278

4 authors:

Md Toki Tahmid
Bangladesh University of Engineering and Technology
20 PUBLICATIONS **84** CITATIONS
SEE PROFILE

Muhammad Ehsanul Kader
Bangladesh University of Engineering and Technology
4 PUBLICATIONS **73** CITATIONS
SEE PROFILE

Tanvir Mahmud
University of Texas at Austin
29 PUBLICATIONS **938** CITATIONS
SEE PROFILE

Shaikh Anowarul Fattah
Princeton University
205 PUBLICATIONS **2,808** CITATIONS
SEE PROFILE

MD-CardioNet: A Multi-Dimensional Deep Neural Network for Cardiovascular Disease Diagnosis from Electrocardiogram

Md Toki Tahmid, *Student Member, IEEE,* Muhammad Ehsanul Kader, *Student Member, IEEE,* Tanvir Mahmud, *Graduate Student Member, IEEE,* and Shaikh Anowarul Fattah, *Senior Member, IEEE*

***Abstract*—Automated classification of cardiovascular diseases from electrocardiogram (ECG) signals using deep learning has gained significant interest due to its wide range of applications. However, existing deep learning approaches often overlook inter-channel shared information or lose time-sequence dependent information when considering 1D and 2D ECG representations, respectively. Moreover, besides considering spatial dimension, it is necessary to understand the context of the signals from a global feature space. We propose MD-CardioNet, an efficient deep learning architecture that captures temporal, spatial, and volumetric features from multi-lead ECG signals using multidimensional (1D, 2D, and 3D) convolutions to address these challenges. Sequential feature extractors capture time-dependent information, while a 2D convolution is applied to form an image representation from the multi-channel ECG signal, extracting inter-channel features. Additionally, a volumetric feature extraction network is designed to incorporate intra-channel, inter-channel, and inter-filter global space information. To reduce computational complexity, we introduce a practical knowledge distillation framework that reduces the number of trainable parameters by up to eight times (from 4,304,910 parameters to 94,842 parameters) while maintaining satisfactory performance compatible with the other existing approaches. The proposed architecture is evaluated on a large publicly available dataset containing ECG signals from over 10,000 patients, achieving an accuracy of 97.3% in classifying six heartbeat rhythms. Our results surpass the performance of some state-of-the-art approaches. This paper presents a novel deep-learning approach for ECG classification that addresses the limitations of existing methods. The experimental results highlight the robustness and accuracy of MD-CardioNet in cardiovascular disease classification, offering valuable insights for future research in this field.**

***Index Terms*—Deep Learning, ECG Signal Analysis, Computer Aided Analysis, Convolutional Neural Network, Computational Complexity**

I. Introduction

Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels that were behind the death of an estimated 19.7 million people worldwide in 2019 [1]. Electrocardiography (ECG) is the most commonly used non-invasive tool to detect CVDs. It captures the heart's electrical activities over time using multiple electrodes attached to the body's surface. In many circumstances, ECG recordings taken over several hours or even days must be evaluated by a cardiologist, which is a time-consuming and challenging task [2]. Moreover, manual assessment of ECG signals may result in subjective interpretation and inter-observer biases [3]. Automated ECG signal analysis and disease detection have gained significant attention from researchers to overcome this issue. It involves several steps, including signal preprocessing, feature extraction, feature selection/reduction, and classification [4]. Machine learning techniques have long been used to evaluate ECG signals [5]–[9]. Discriminant analysis-based classification of ECG signals is widely used as an alternative to many other machine-learning techniques. In [10], authors have proposed a new scheme based on tensor rank one discriminant analysis in which ECG signals are represented with third-order tensors having spatial, temporal, and spectral domains. Tensor discriminant based analysis with partial labeling is presented in [11].

The above-mentioned algorithmic approaches often require manual feature engineering for better accuracy. Here, deep learning comes into play with its automated feature extraction ability. In [12], an artificial neural network is proposed for Arrhythmia classification using ECG signals. Using a deep neural network (DNN) with a stacked restricted Boltzmann machine (RBM), a beat-by-beat ECG classification technique is proposed in [13]. However, such multilayer perceptron (MLP) based DNNs do not include sequential time series features in the generated feature vectors. The recurrent neural network (RNN) is a better choice as ECG is a time series [14]. A bidirectional long short-term memory (LSTM) model is utilized in [15] to classify 12 rhythm classes. A new model for deep bidirectional LSTM network-based wavelet sequences called DBLSTM-WS is proposed in [16] for classifying ECG signals. Using solely sequence models to extract information from time-series data may lose crucial features from inter-channel space.

Convolutional neural networks (CNNs) have recently been proven effective when dealing with time-series data with multiple channel features [17]–[20]. A 16-layer 1D CNN is proposed in [21] to classify Arrhythmia. DeepArrNet architecture is proposed in [22] for automatically detecting Arrhythmia in which a series of different structural units are utilized. Nonetheless, 1D convolution, similar to sequential RNN, fails to extract inter-channel information from spatial feature space. For spatial feature extraction, recently in [23], images are formed from the multi-channel ECG database, used to train a 2D CNN model, and found a very satisfactory performance.

T. Mahmud and S. A. Fattah are with the Department of EEE and MT Tahmid and ME Kader are with and CSE Department, Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh e-mail: (tanvirmahmud@eee.buet.ac.bd, fattah@eee.buet.ac.bd, sharifulislam-toki@gmail.com, and ehsanul.kader16@gmail.com).

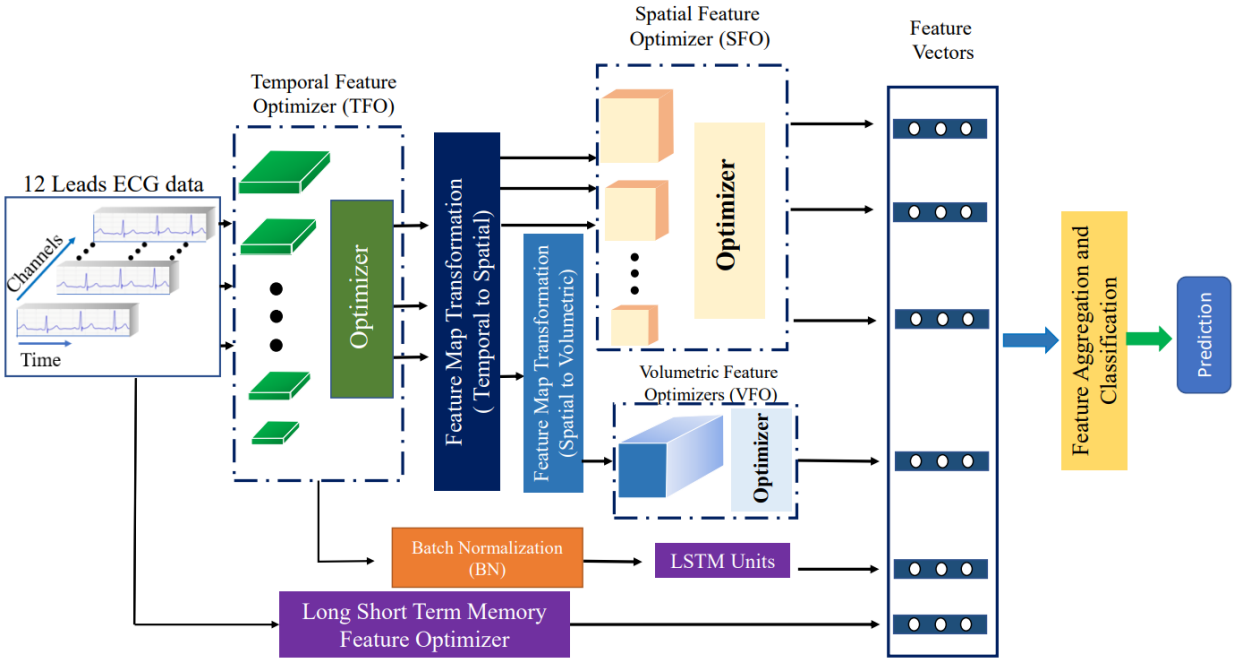


Fig. 1: **Workflow of the proposed scheme for diagnosis of cardiac disease from 12 lead ECG signals. Firstly, temporal and spatial feature maps are extracted using temporal and spatial feature optimizers. Volumetric feature optimizer extracts information from a 3D feature space. Two LSTM optimizers are used for better conservation of sequential features. Finally, feature vectors are aggregated for making predictions**

All the above-proposed approaches face the issue of increasing computational complexity when trained with more complex architectures. Moreover, the interconnection between different leads while dealing with feature extraction from multi-lead ECG data is also ignored in the previous literature Applying 1D, 2D, and even 3D CNN together may provide a superior result for detecting diseases from ECG data yet to be examined.

This paper proposes a multidimensional (1D, 2D, and 3D CNN-based) approach for classifying diseases from multi-lead ECG signals and a knowledge distillation based schema to reduce the computational complexity. We try to ensure the model's robustness against various input and initial conditions by cross-fold validation and random initialization. Moreover, the applicability of the scheme in real world is ensured by reducing the running time significantly by using knowledge distillation. The significant contributions of this work are summarized below:

- In the proposed architecture, along with sequential feature extractors to capture time-dependent information, an image representation is formed using the given multi-channel ECG signal, and 2D convolution is carried out to extract inter-channel features. Moreover, a volumetric feature extraction network is designed, which provides information from intra-channel, inter-channel, and inter-filter global space.
  This has shown significant improvement in accuracy, and the placement of the volumetric feature extractor is scrutinized so that the model does not face an over-fitting issue. Such incorporation of 2D mapping and volumetric feature optimization for ECG signal analysis is the first of its own. We have achieved State-Of-The-Art performance over the studied benchmark dataset.
- We propose a novel and effective knowledge distillation framework that distills the knowledge from the best-performed teacher model to a reduced student model by a custom loss function that reduces the cost of feature vectors. The knowledge distillation network uses the same architecture as the best-performed proposed model with the reduction in parameter number in each Layer. By expanding the traditional knowledge distillation framework, we have distilled knowledge to the reduced model from the four major components (temporal optimizer, spatial optimizer, volumetric optimizer, LSTM optimizer). This allows us to reduce the model size to 32 times without losing much accuracy.

II. Methodology

This section presents the fundamental concept behind the proposed architecture and the training scheme. Next, the detailed description, along with the necessity of various sub-modules of the proposed MD-CardioNet, is discussed in the following subsections. Finally, the proposed effective knowledge distillation strategy for reducing computational complexity is presented. Also, the dataset description and data sampling methods are provided.

A. Dataset Description

Experiments are carried out on a publicly available dataset containing 12 lead ECG records from 10,464 patients, including 5,956 males and 4,690 females, to validate the proposed architecture's effectiveness and the sophisticated knowledge

TABLE I: **DESCRIPTION OF THE ARRHYTHMIA CLASSES**

Name	Frequency, n(%)	Females	Males
SB (Sinus Bradycardia)	3,889(36.53)	1408	2481
SR (Sinus Rhythm)	1,826(17.15)	1024	802
AFIB (Atrial Fibrillation)	1,780(16.72)	739	1041
SI (Sinus Irregularity)	399(3.75)	175	224
AF (Atrial Flutter)	445(4.18)	182	263
TCD (Tachycardia classes)	2300(23)	1135	1165

distillation algorithm. Chapman University and Shaoxing People's Hospital collected the ECG dataset used in his paper. The dataset contains 12 lead ECG records from 10,464 patients, allowing us to examine the validity of any proposed scheme rigorously. In the original dataset, Patients are categorized into 11 arrhythmia classes, namely, Sinus Bradycardia (SB), Sinus Rhythm (SR), Atrial Fibrillation (AFIB), Sinus Tachycardia (IST), Atrial Flutter(AF), Sinus Irregularity(SI), Superventricular Tachycardia(SVT), Atrial Tachycardia(AT), Atrioventricular Reentrant Tachycardia (AVRT), Atrioventricular Node Reentrant Tachycardia (AVNRT), and Sinus Atrium to Atrial Wandering Rhythm (SAAWR).

Like most references, we consider six classes, with five different classes: SB, SR, AFIB, AF, SI, and a combined class of Tachycardia classes (TCD).

Arrhythmia classes and their distributions in this dataset are presented in Fig.I. Due to insufficient records (only 0.03%), the SAAWR class is not included in the final dataset for classification. Classes containing fewer data points are upsampled to make the dataset distribution symmetric, with 2000 records per class. We have used two approaches for the data augmentation task. Firstly, we vary the given signals using translation. Secondly, we add random noise to the signal. During translation, we shift the signal randomly along the x and y axis; this cuts some parts of the actual signal, increasing the model's ability to learn from a comparatively challenging feature space. For introducing random noise, we first normalize the whole signal, and then for randomly chosen time stamps, we add a random normalized value to the actual signal value for that time stamp. Thus, the final processed dataset contains six classes: SB, SR, AFIB, AF, SI, and TCD.

Each record is 10.24 seconds long with a sampling frequency of 100Hz. Thus each record consists of 1024 data points. 1024 data points are taken for efficient scaling while training different network layers. .

B. Proposed Objective Function with Multidimensional Convolutional Operations

In Fig.1, the basic blocks involved in the proposed architecture are shown. In order to utilize 1D, 2D, and 3D CNN-based spatiotemporal operations on the given multi-lead EEG data, in the proposed architecture, we introduce three optimizers, namely temporal feature optimizer (TFO), spatial feature optimizer (SFO), and volumetric feature optimizer (VFO). Moreover, an LSTM-based optimizer is also employed. After developing the base model using this architecture, we scale the network parameters to a great extent to reduce computational complexity using a custom knowledge distillation network.

In the training stage, the 12 lead ECG signals and their corresponding labels (diseased or healthy classes) are passed through the network to learn their weights and biases and get an optimized feature vector at each sub-network. Finally, feature vectors from these branches are concatenated to obtain a final probabilistic distribution.

Let us consider the set of 12-lead ECG signals as $\mathbf{X}$, and their corresponding ground truths as $\mathbf{Y}$, where $X_i \in \mathbb{R}^{t\times c}$, $Y_i \in \mathbb{R}^{L}$, and i=$1, 2, 3, ...N$. Here t denotes the length of the ECG signal, c denotes the number of leads, L denotes the number of target labels (here, the number of predicted cardiac diseases), and N is the total number of ECG records. Given any particular ECG record $X_i \in \mathbb{R}^{t\times c}$, the target is to predict the label of this record among all possible labels. As shown in Fig.1, multidimensional feature extraction and training are proposed in our network, including temporal, spatial, and volumetric feature optimization. First, in the TFO stage of the network, the objective function optimizes the temporal feature maps. This unit is composed of one-dimensional convolution and pooling layers.

Next, in the SFO step, the temporal feature maps generated by the TFO are transformed into spatial feature maps and then manipulated by the SFO unit. In the third step, spatial feature maps are primarily transformed into volumetric features and then manipulated using the VFO. The method used to convert feature map dimensions is presented in Fig.2. The selected units of the spatial feature maps for avoiding over-fitting and extracting the most significant volumetric features will be discussed in the following sections. In addition to the multi-dimensional convolution, a parallel LSTM branch is operated over 12 leads simultaneously, denoted as an LSTM feature optimizer (LTSM-FO). This module extracts time-dependent sequential features rather than capturing localized information like the convolutional network.

In the objective function, we want to minimize three types of loss functions $L\{.\}$, namely (1) temporal loss function L_t, (2) spatiotemporal loss function L_{st} and (3) volumetric-spatial-temporal loss function L_{vst}. Thus the overall objective function of the described branches is formulated as follows:

$$
\begin{aligned}
Ls &= L_t + L_{st} + L_{vst} \\
L_t &= \underset{\theta_t}{\operatorname{argmin}} \mathscr{L}(\theta_t, y_t^p, y_t) \\
L_{ts} &= \underset{\theta_t,\theta_s}{\operatorname{argmin}} \{\mathscr{L}(\theta_s, y_t^p, y_t), \mathscr{L}(\theta_s, y_s^p, y_s)\} \\
L_{vst} &= \underset{\theta_t,\theta_s,\theta_v}{\operatorname{argmin}} \{\mathscr{L}(\theta_t, y_t^p, y_t), \mathscr{L}(\theta_s, y_s^p, y_s), \mathscr{L}(\theta_v, {y_v}^p, y_v)\}
\end{aligned} \tag{1}
$$

Here θ_t denotes the temporal network parameters, y_t^p denotes the predicted temporal probability mask, and y_t depicts the ground temporal probability mask for the sequential feature optimizer branch. Similarly, θ_t,y_t^p, and y_t denote spatial network parameters, predicted spatial probability mask, and ground probability mask respectively. Parameters sub-scripted with v denote the same for the volumetric network. Here, by

probability mask, we denote the prediction scores obtained from a particular layer within the network when input is fed into it. The predicted probability mask is the prediction score of a layer when an input sample is run through the layers. The ground probability mask (ground truth) is the prediction score that the Layer receives during back-propagation. We try to minimize the loss between these two probability masks for each Layer. Similarly, subscripts s and t denote the branches for spatial and volumetric feature optimizers.

We perform a fusion of multidimensional convolutional operations in the proposed cardiovascular disease analysis network. Generally, when a sequential network is applied to time-series data, it loses information about the inter-channel shared features. We have mapped the sequential feature optimizers into a 2D spatial matrix and operated 2D convolution to capture inter-channel features to overcome this issue. Finally, we thoroughly examined the possibility of incorporating an effective volumetric convolution to get features from a global space incorporating intra-channel, inter-channel, and inter-filter components.

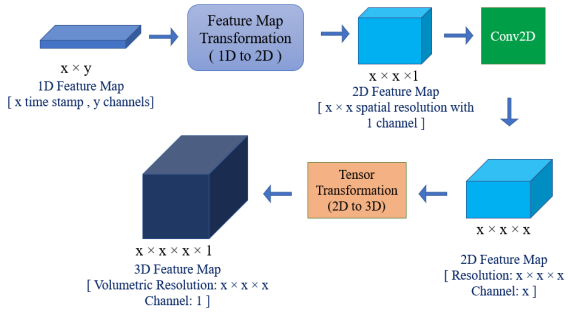


Fig. 2: **Schemetic representation of the feature map transformation network. At first, the 1D feature map is transformed into a 2D map by extending the channel dimension. The resolution of the 2D map is extended to 3^{rd} dimension followed by 3D convolution**

C. Proposed Sequential Feature Extractors

In order to preserve the time-dependent sequential information and retrieve features simultaneously from the 12 channels of the ECG signal, two parallel sequential feature extractors, namely TFO and LSTM-FO, are operated.

1) Temporal Feature Optimizer (TFO) : Input to this layer is the ECG signal X, such that $X_i \in \mathbb{R}^{t \times c}$. This sub-unit uses a series of 1D-Net layers to extract sequential and inter-channel features. The architecture of the TFO unit is shown in Fig.3. Each 1D-Net Layer consists of one conv1D unit and a max-pooling unit. Here, the pooling size is fixed at 4, and the number of filters is doubled in each unit. The number of 1D-Net units determines the depth of this module. If the number of 1D-Net units is L, this network is represented as

$$D_1(a,b,c) \rightarrow ... \rightarrow D_L(\frac{a}{4^{L-1}}, \frac{b}{4^{L-1}}, c \times 2^{L-1}) \quad (2)$$

where each D_i in i= 1,2,...L denotes 1D-Net unit.

TABLE II: **ARCHITECTURE OF THE SPATIAL FEATURE OPTIMIZER UNIT (F= FILTER NUMBER, K= KERNEL, S= STRIDE, N= SINGLE NEURON NODE). FOR POOLING AND FULLY CONNECTED LAYERS, NO FILTER AND STRIDES ARE APPLIED. SINGLE NEURON NODE IS ONLY APPLICABLE FOR FULLY CONNECTED LAYER. HERE A= HEIGHT, B=WIDTH, AND C=NUMBER OF CHANNELS IN THE INPUT ECG SIGNAL**

Input Size	Operation	f	K	S	N	Output Size
(a,b,c)	Conv2D	32	14,14	1,1	-	(a,b,32)
(a,b,32)	MaxPool2D	-	2,2	-	-	(a/2,b/2,32)
(a/2,b/2,32)	Conv2D	64	7,7	1,1	-	(a/2,b/2,64)
(a/2,b/2,64)	MaxPool2D	-	2,2	-	-	(a/4,b/4,64)
(a/4,b/4,64)	Conv2D	128	5,5	1,1	-	(a/4,b/4, 128)
(a/4,b/4, 128)	MaxPool2D	-	2,2	-	-	(a/8,b/8, 128)
(a/8,b/8, 128)	Global AvgPool2D	-	a/8, b/8	-	-	128
128	Fully Connected	-	-	-	512	512

a and b represent the length and width of the 1D feature map, with c represents the number of filters. At the end of this network, to extract sequential information from the local optimizers, global average pooling is applied, followed by a layer of LSTMs. The intuition behind using the LSTM layer after temporal feature extraction is to ensure that time-dependent sequential information is not lost during the convolutions. The addition of this Layer also provides us with improvement in the overall performance.

2) Multilayer LSTM Feature Optimizer: This sub-network comprises Bidirectional Long Short Term Memory (LSTM) layers. Before passing the input in this Layer, BatchNormalization (BN) is applied over the input, allowing LSTM networks to converge fast.

Bidirectional LSTM networks allow the extraction of information from both directions of the time axis. This module analyzes sequential features by integrating all the leads simultaneously, ensuring the conservation of sequential time dependence.

D. Spatial Feature Optimizer (SFO)

The spatial feature optimizer (SFO) unit introduced in the proposed architecture (shown in Fig.1) employs multidimensional convolution to capture spatial features from the ECG signal. As shown in Fig.3, square temporal feature maps of dimension $x \times x$ are extracted from the primary sequential feature network. With one channel extended resolution, these maps are transformed into 2D feature maps. These 2D feature maps are passed through the SFO unit to optimize inter-channel spatial relations. The architecture of the SFO unit is presented in Fig.II. Here, three consecutive Conv2D and MaxPool2D layers are used with pooling size 2X2, and the number of filters doubled in each convolution. After operating the convolutions, a Global Average pooling layer is used across all the channels and finally passed through a dense layer containing 512 neurons.

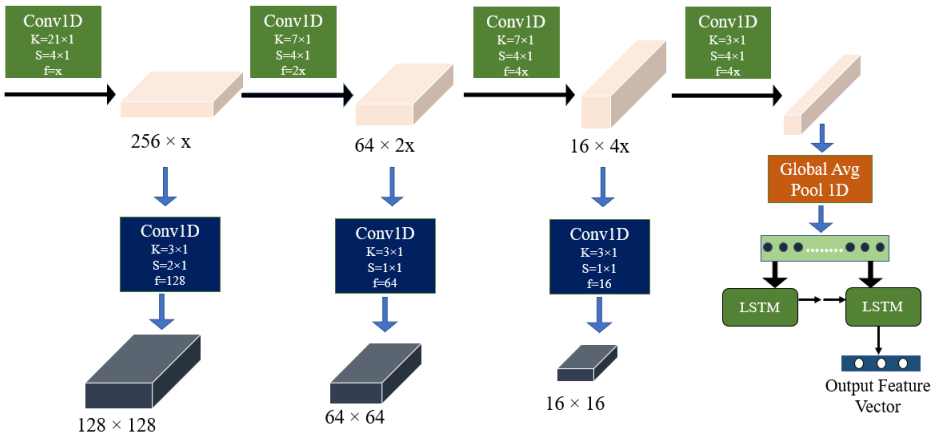


Fig. 3: **Schematic representation of the primary temporal feature extraction network. In the TFO of each 1D-Net unit, the resolution size is reduced by 4 times, and the channel size is doubled. Here, we extract three secondary branch networks and transform them into a square 1D feature vector manipulated by the SFO unit in the next part. .**

TABLE III: **ARCHITECTURE OF THE VOLUMETRIC FEATURE OPTIMIZER UNIT (F= FILTER NUMBER, K= KERNEL, S= STRIDE, N= SINGLE NEURON NODE). FOR POOLING AND FULLY CONNECTED LAYERS, NO FILTER AND STRIDES ARE APPLIED. SINGLE NEURON NODE IS ONLY APPLICABLE FOR FULLY CONNECTED LAYER. HERE A= HEIGHT, B=WIDTH, C=NUMBER OF CHANNELS, AND D=NUMBER OF VOLUMETRIC UNITS IN THE INPUT ECG SIGNAL**

Input Size	Operation	f	K	S	N	Output Size
(a,b, c,d)	Conv 3D	32	14,14	1,1	-	(a,b, c,32)
(a,b, c,32)	MaxPool 3D	-	2,2	-	-	(a/2,b/2, c/2,32)
(a/2,b/2, c/2,32)	Conv 3D	64	7,7	1,1	-	(a/2,b/2, c/2,64)
(a/2,b/2, c/2,64)	MaxPool 3D	-	2,2	-	-	(a/4,b/4, c/4,64)
(a/4,b/4, c/4,64)	Conv 3D	128	5,5	1,1	-	(a/4,b/4, c/4,128)
(a/4,b/4, c/4,128)	MaxPool 3D	-	2,2	-	-	(a/8,b/8, c/8,128)
(a/8,b/8, c/8,128)	Global AvgPool3D	-	a/8, b/8, c/8	-	-	128
128	Fully Connected	-	-	-	512	512

E. Volumetric Feature Optimizer (VFO)

We investigated if further expansion in dimensional features helps the model get a global view, including the inter-channel, intra-channel, and inter-filter space. Although leveraging all L numbers of the 2D feature maps by extending their resolution in 3D helps the model extract more sophisticated features, it reduces the model's generalization ability. Hence, the VFO is placed at the last Layer of the sequential feature optimizer network, which contains the most crucial spatial features. For this, if the initial dimension of the temporal feature map is $\mathbb{R}^{x\times y}$, first it is converted into a spatial feature map of dimension $\mathbb{R}^{x\times x\times 1}$ by extending the resolution of the feature map. On this transformed map, a 2D convolution is applied with x filters to get a shape of $\mathbb{R}^{x\times x\times x}$. Similarly, we perform the volumetric transformation, and the final shape of the feature map becomes $\mathbb{R}^{x\times x\times x\times 1}$. The architecture of the VFO unit is presented in Fig.III. Similar to the SFO network, in the VFO unit, consecutive Conv3D and MaxPool3D operations are used. The initial input size is halved in each step, doubling the number of filters. Finally, a global average pooling layer followed by a dense layer of 512 neurons is used to produce the resulting volumetric feature map.

F. Proposed Knowledge Distillation Based Sequential Training Algorithm

Reducing dimension and computational complexity remains pivotal while deploying deep learning models into real-time production. Knowledge distillation has gained substantial attention in this regard [24]. In this technique, a large model, namely the teacher model, is trained at first. Then the masked probability values of the edge layers of this network are distilled using a comparatively small student model. In the classic case, the loss function of the student model is defined as:

$$L_{student} = \alpha * KL\left(softmax(\frac{\vartheta_T}{\Gamma}), softmax(\frac{\vartheta_S}{\Gamma})\right) + (1-\alpha) * crossentropy\_loss\ (Y, Y^P) \quad (3)$$

Here, α is the weight given to the distillation loss. ϑ_T denotes the last prediction layer for the teacher model, and ϑ_S denotes the last prediction layer for the student model. We take the KL divergence between these two layers and optimize the error. We use categorical cross-entropy to calculate the loss between the predicted and true classes along with the distillation loss. Both these losses are added and optimized through back-propagation. The distributions are softened by applying a temperature scaling factor denoted as Γ, smoothing out the probability distribution. Here, α is the weight given to the distillation layers, and $1-\alpha$ is the weight given to the final prediction layer. Y and Y^P denote the prediction and true label scores, respectively.

In the proposed method, this loss function is customized to incorporate and distill the knowledge from feature vectors

corresponding to the teacher and student models' temporal, spatial, volumetric, and sequential optimizer branches. As shown in Fig.4, the teacher model is modified and retrained by transforming the final feature vectors into a dimension-compatible reduced-sized student model. Then, rather than taking only the prediction layer, we take all the reduced feature layers of the teacher model to distill the knowledge to the student model. Thus the loss function for the proposed knowledge distillation network is expressed as:

$$\mathscr{L} = \varphi_1\ CrossEntropy(Y, Y^p) + \varphi_2\ KL_{div}(f_1, f_1') + \varphi_3\ KL_{div}(f_2, f_2') + \varphi_4\ KL_{div}(f_3, f_3') + \varphi_5\ KL_{div}(f_4, f_4') \tag{4}$$

Here, φ_i denotes the weight assigned to each distillation and prediction layer and $\sum_{i=1}^{5} \varphi_i = 1$. f_1 denotes the reduced feature vector of the temporal optimizer from the teacher model, and f_1' denotes the reduced feature vector of the temporal optimizer from the student model. Subscripts 2, 3, and 4 denote the same for spatial, volumetric, and sequential LSTM feature optimizers. The fully connected Layer is manipulated to correspond to the dimensions of the reduced-size student model.

$Crossentropy$ loss function is applied to the student model's prediction layer. As shown if Fig.4, the total loss is calculated as the sum of the distillation loss and the predicted loss with hard labels. Here by the hard label, we denote the corresponding output class of the input signal. The loss is back-propagated to optimize the network.

III. Results and Discussions

For this section, the classification performance of the proposed scheme is evaluated on a widely used public ECG dataset. Performances of the various sub-modules of the network are studied separately to better understand the contributions of each dimension of the proposed multidimensional framework.

A. Experimental Setup

Computations are performed on the Google Colab platform for accelerated performance with GPU support and 12 GB virtual ram for training deep learning models.

The models are developed with Tensorflow. For training and validation of the proposed model, we use the Keras framework from Tensorflow. We train the model with 50 epochs and a batch size of 512. In each epoch, 512 samples are trained through back-propagation, and each of these epochs takes roughly 35s to run. We use Adam optimizer and a learning rate of 0.001 in this study.

The computational complexity of the proposed deep neural network is computed in terms of the number of trainable parameters and inference time. For the proposed best-performed model, the number of parameters is 4,304,910, with an inference time of 0.5s.

Several traditional evaluation metrics are used for the evaluation of performance. These are given by:

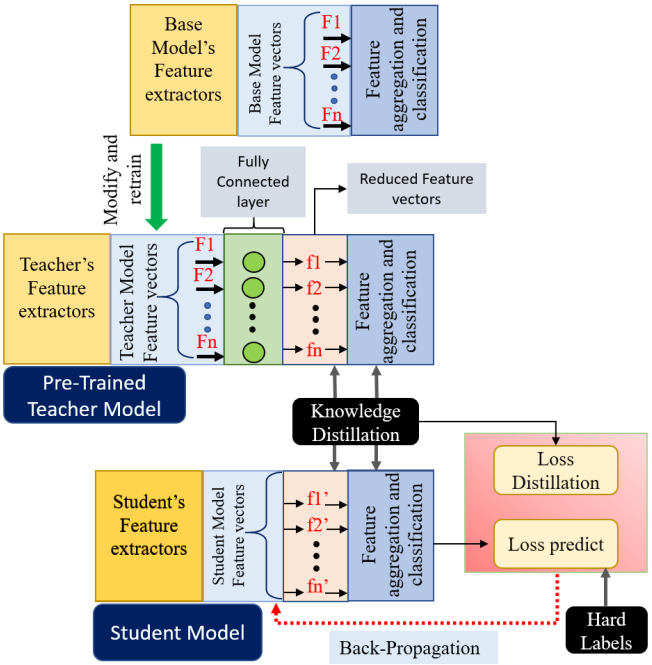


Fig. 4: **Schematic representation of the proposed knowledge distillation network. The best-performed base model is modified and retrained by manipulating the fully connected Layer to correspond with the student model. The combined loss function is calculated using the reduced feature layers and the Hard Labels loss. Effective back-propagation ensures the optimization of the Student Model.**

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1\ Score = \frac{2 * Precision * Recall}{accuracy + Recall} \tag{8}$$

Here, TP refers to the number of samples predicted to be in a particular class, and they reside in that class. FP refers to samples predicted to be in a particular class, but in the test dataset, they do not belong to that class. TN refers to the samples predicted not to be a sample of a disease class that does not belong to that class. Finally, FN is samples that are predicted as not being a class member when they belong to that class in a given dataset.

B. Performance Evaluation and Ablation Study

With the Md-CardioNet architecture, we have achieved an accuracy of 97.3% on the studied dataset. A hold-out validation scheme is used here, as done in previous literature. We also perform a ten-fold cross-validation test on the whole dataset to validate the model's performance with different validation sets. With the ten-fold cross-validation method, the

TABLE IV: **CLASS-WISE ACCURACY, RECALL, F1-SCORE AND SUPPORT. OVERALL ACCURACY, MACRO AND MICRO AVERAGE VALUES OF OTHER METRICS ARE PROVIDED. (HERE, SUPPORT=NUMBER OF SAMPLES USED)**

	Accuracy	Recall	F1-Score
SB	0.99	0.96	0.97
SR	0.98	0.96	0.97
Tachycardia	0.98	0.96	0.97
AFIB	0.97	0.97	0.97
AF	0.97	1.00	0.99
SI	0.96	1.00	0.98
Accuracy			0.97
Macro average	0.98	0.97	0.97
Weighted average	0.98	0.97	0.97

TABLE V: **ABLATION STUDY OF THE EFFECT OF DIFFERENT MODULES IN THE PERFORMANCE OF THE PROPOSED MD-CARDIONET ARCHITECTURE**

Model	Module Placement	Accuracy
Baseline(1D CNN)	Series 1D CNN	0.857
Baseline + LSTM	Parallel LSTM	0.874
	Series LSTM	0.892
Baseline + LSTM + 2D Net	Branch: 1	0.915
	Branch: 1 and 2	0.939
	Branch: 1, 2, and 3	0.942
Baseline+ LSTM + 2D Net + 3D Net (MD-CardioNet)	Branch: 1	**0.973**
	Branch: 1 and 2	0.955
	Branch: 1, 2, and 3	0.962

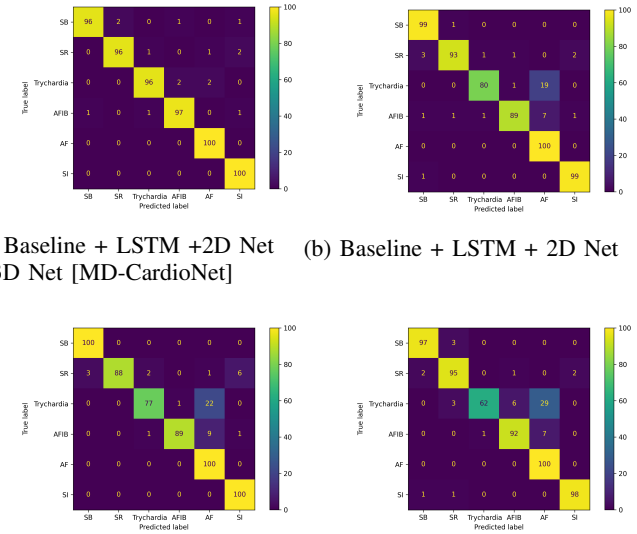
validation accuracy is found as 97.04%, which indicates the compatibility of the proposed architecture with any validation dataset within the given data. The standard deviation among the folds is 0.49%. We confirm that the samples used as train data are not present in the validation part, because the 10,000 samples we use, are from 10,000 different persons. To better understand how well the model classifies each class individually, class-wise performance metrics are provided in Fig.IV.

An ablation study is carried out to analyze the effectiveness of different modules of the proposed MD-CardioNet architecture. The baseline model is constructed using a series 1D CNN model, and various sub-modules of the proposed architecture are added to validate performance improvement.

1) Effect of the LSTM Sub-Network: The baseline model with series 1D CNN provides an accuracy of 85.7% as presented in Fig.V. Along with the 1D CNN, a parallel LSTM network is introduced, as shown in the LSTM Optimizer unit of Fig.1. This module increases the accuracy by 1.7%.

TABLE VI: **DESCRIPTION OF MODEL PARAMETERS AND RESULTS BASED ON DIFFERENT SCALING**

Scaling Factor	Parameters	Accuracy	Recall	F1 Score	Inference Time
1	4,304,910	**0.973**	**0.962**	**0.966**	0.501s
2	1,138,518	0.960	0.952	0.951	0.432s
4	315,886	0.942	0.933	0.929	0.331s
8	94,842	0.921	0.914	0.907	0.310s
16	31,888	0.913	0.891	0.899	0.209s
32	20,000	0.875	0.854	0.866	**0.182s**



(a) Baseline + LSTM +2D Net + 3D Net [MD-CardioNet]

(b) Baseline + LSTM + 2D Net

(c) Baseline + LSTM

(d) Baseline Model

Fig. 5: **Confusion metrics of the ablation study.**

TABLE VII: **PERFORMANCE OF REDUCED MODELS ON DIFFERENT DISTILLATION SCHEMES**

Model	Parameters	accuracy		
		Without Distillation	With Last layer distillation	With Proposed Distillation
Teacher Model	4,304,910	0.973	-	-
Student Model (16 times scaling)	31,888	0.913	0.933	**0.96**
Student Model (32 times scaling)	20000	0.875	0.912	**0.945**

Along with the parallel LSTM network, a series LSTM optimizer is introduced along with batch normalization after the TFO network, which provides a further 1.8% improvement in accuracy.

2) Effect of the 2D Sub-Networks: The 2D feature optimizer CNN block helps to improve the accuracy to a significant extent. Fig.V shows the effect of changing the number of 2D branches on accuracy. With only one branch of the 2D network, the best performance achieved is 91.5%, a 2.3% improvement over the Baseline + LSTM network. While keeping the validation and test scores into proper consideration, the best performance is achieved with three branches of the 2D network, which is 94.2%.

3) Effect of 3D-Sub Network: For volumetric feature extraction, 3D-Sub Networks are introduced carefully, which increases the model complexity. However, it is found that including multiple 3D branches over the network could provide more improvement; rather often faces overfitting. As shown in Fig.V, extending the network with only one 3D module at the last unit of the 2D Sub-Network, helps the model learn the best possible features from the given dataset. With this setup, our model provides an accuracy of 97.3%, which is an 11.6%

TABLE VIII: **PERFORMANCE COMPARISON WITH SOME OTHER EXISTING METHODS. ALL THESE METHODS USE 12 LEAD ECG DATA AS INPUT**

Methods	Dataset	Subjects	Method	Performance
[25]	Chapman University and Shaoxing People's Hospital	10,646	DNN	92.24%
[26]	Chapman University and Shaoxing People's Hospital	10,646	HIT+ SVM	92.95%
[27]	Chapman University and Shaoxing People's Hospital	10,646	Transfer learning	87%
Proposed	Chapman University and Shaoxing People's Hospital	10,646	CNN+ LSTM	**97.3%**

improvement from the baseline model and the best accuracy on this large dataset until now.

The confusion metrics of the ablation study and performance improvement with incorporated modules are shown in Fig.5.

C. Effect of the Distillation Operation on Computation

In order to demonstrate the effect of the proposed knowledge distillation algorithm in reducing the computational complexity as well as the inference time, an experiment is conducted by scaling the parameters of the best-performed model.

As shown in Fig.VI, we have scaled the network parameters while keeping the baseline architecture and investigating its performance. Reducing the parameter size significantly improves the inference time but drops the accuracy to a large extent when scaled 16 or 32 times.

At this point, we have used our proposed distillation network. The best-performed model is taken as the teacher model. The scaled networks are trained as student models with the traditional knowledge distillation of the last distribution layer and the proposed multilayer knowledge distillation. As presented in Fig.VII, last layer distillation improves the accuracy of the 16 times and 32 times scaled models by 2% and 3.7%, respectively. At the same time, the proposed distillation model improves this accuracy by 4.7% and 7%.

D. Comparison with Other Existing Approaches

Several state-of-the-art networks are considered to compare the performances of the proposed Md-CardioNet. Three methods reported in Fig.VIII considered the same dataset used in this paper. The method proposed in [25] uses a stacked DNN and a one-dimensional CNN-based approach to classify heart disease. Here the interaction between different ECG leads needs to be considered, which lacks spatial information. In [26], authors have proposed a method to extract spatial information using a homomorphically irreducible tree(HIT) and perform classification using a support vector machine (SVM).In such cases, the temporal features are ignored. Moreover, SVM-based methods face difficulties when dealing with challenging classes, as the decision boundary of SVM often fails to capture complex internal relations between classes. In [27], a transfer-learning-based method and Bootstrapping are proposed. The above-proposed approaches simultaneously use temporal and spatial features while dealing with multi-lead ECG data. The shortcomings of the previously proposed approaches are addressed using MD-CardioNet. The issues with typical SVM-based classifiers, as used in [26], is not present in the proposed architecture as it simultaneously captures information from all three dimensions making it possible to build complex decision boundary with challenging classes. Instead of using one-dimensional feature space only, as in [25], we employ spatial feature space into the architecture that helps to capture the correlation between different ECG leads. Moreover, instead of using transfer-learning as done in [27], we train the architecture from scratch and reduce its dimensionality using knowledge distillation, resulting in faster inference time with satisfying accuracy.

E. Limitations and Future Studies

The study is conducted on a single source large public dataset. As a future work, MD-CardioNet could be implemented on other custom datasets not only restricted to ECG signals. Moreover, multi-modal bio-signal analysis [28], [29] can be explored by leveraging the different layers of the multidimensional network model. Parts of the ECG signal that contribute most to making a particular prediction can be explored using gradient-based approaches, such as the grad-cam method.

Real-time implementation and deployment are not performed in the current study. The recent parallel and cloud computing development allows complex network models to operate without centralized sophisticated hardware support. In [30], [31], cloud computing-based remote health monitoring schemes are proposed based on single-channel ECG data, which can be extended to multi-channel ECG monitoring system deployment.

IV. Conclusion

This study proposes an automated scheme using an efficient neural network architecture (MD-CardioNet) to classify cardiovascular diseases based on 12 lead ECG signals. The major contribution of the proposed architecture is to capture the temporal, spatial, and volumetric features very effectively for getting inter-channel and intra-channel crucial information among the leads of the ECG signal.

We show in the ablation study that the use of feature map transformation from 1D to 2D image space increases the accuracy significantly. Placement of the volumetric feature extractors is studied extensively, which offers further improvement in results without facing the issue of over-fitting. The

accuracy is at least 5.4% higher than that obtained by some existing methods on the studied dataset. Thus to the best of our knowledge, we have introduced a fusion of multidimensional feature extractors to retrieve essential information about multi-channel ECG data for the first time. Another significant contribution of this research includes the incorporation of a novel multi-layer knowledge distillation framework, which distills the knowledge from the teacher model to a student model from the three convolution dimensions and the sequential feature space. This allows for reducing the parameter counts of the model up to 16 times without losing much accuracy.

Therefore, MD-CardioNet can be reliably implemented in identifying cardiovascular diseases from ECG data, even if high computational power is unavailable.

Data and Code Availability

The code and used dataset for MD-CardioNet are provided in the following GitHub repository:
Implementation of MD-CardioNet

References

[1] G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Z. Beaton, E. J. Benjamin, C. P. Benziger, *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990–2019: update from the gbd 2019 study," *Journal of the American College of Cardiology*, vol. 76, no. 25, pp. 2982–3021, 2020.

[2] P. Sodmann, M. Vollmer, N. Nath, and L. Kaderali, "A convolutional neural network for ecg annotation as the basis for classification of cardiac rhythms," *Physiological Measurement*, vol. 39, no. 10, p. 104005, 2018.

[3] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, M. Adam, A. Gertych, and R. San Tan, "A deep convolutional neural network model to classify heartbeats," *Computers in Biology and Medicine*, vol. 89, pp. 389–396, 2017.

[4] A. Y. Hannun, P. Rajpurkar, M. Haghpanahi, G. H. Tison, C. Bourn, M. P. Turakhia, and A. Y. Ng, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65–69, 2019.

[5] R. J. Martis, C. Chakraborty, and A. K. Ray, "Wavelet-based machine learning techniques for ECG signal analysis," in *Machine learning in healthcare informatics*, pp. 25–45, Springer, 2014.

[6] K. Polat, B. Akdemir, and S. Güneş, "Computer aided diagnosis of ECG data on the least square support vector machine," *Digital Signal Processing*, vol. 18, no. 1, pp. 25–32, 2008.

[7] F. Melgani and Y. Bazi, "Classification of electrocardiogram signals with support vector machines and particle swarm optimization," *IEEE Transactions on Information Technology in biomedicine*, vol. 12, no. 5, pp. 667–677, 2008.

[8] N. Ghoggali, F. Melgani, and Y. Bazi, "A multiobjective genetic svm approach for classification problems with limited training samples," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 6, pp. 1707–1718, 2009.

[9] Z. Zidelmal, A. Amirou, D. Ould-Abdeslam, and J. Merckle, "ECG beat classification using a cost sensitive classifier," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 3, pp. 570–577, 2013.

[10] K. Huang and L. Zhang, "Cardiology knowledge free ecg feature extraction using generalized tensor rank one discriminant analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–15, 2014.

[11] K. Huang, D. Zhou, Y. Cong, W. Xu, W. Wang, and Z. Que, "Tensor discriminant analysis with partial label," *Procedia computer science*, vol. 131, pp. 416–424, 2018.

[12] G. Sannino and G. De Pietro, "A deep learning approach for ECG-based heartbeat classification for arrhythmia detection," *Future Generation Computer Systems*, vol. 86, pp. 446–455, 2018.

[13] S. S. Xu, M.-W. Mak, and C.-C. Cheung, "Towards end-to-end ECG classification with raw signal extraction and deep neural networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1574–1584, 2018.

[14] M. Hüsken and P. Stagge, "Recurrent neural networks for time series classification," *Neurocomputing*, vol. 50, pp. 223–235, 2003.

[15] K.-C. Chang, P.-H. Hsieh, M.-Y. Wu, Y.-C. Wang, J.-Y. Chen, F.-J. Tsai, E. S. Shih, M.-J. Hwang, and T.-C. Huang, "Usefulness of machine learning-based detection and classification of cardiac arrhythmias with 12-lead electrocardiograms," *Canadian Journal of Cardiology*, vol. 37, no. 1, pp. 94–104, 2021.

[16] Ö. Yildirim, "A novel wavelet sequence based on deep bidirectional lstm network model for ECG signal classification," *Computers in Biology and Medicine*, vol. 96, pp. 189–202, 2018.

[17] J. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[18] K. N. Nabi, M. T. Tahmid, A. Rafi, M. E. Kader, and M. A. Haider, "Forecasting COVID-19 cases: A comparative analysis between recurrent and convolutional neural networks," *Results in Physics*, vol. 24, p. 104137, 2021.

[19] B. Zhao, H. Lu, S. Chen, J. Liu, and D. Wu, "Convolutional neural networks for time series classification," *Journal of Systems Engineering and Electronics*, vol. 28, no. 1, pp. 162–169, 2017.

[20] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-d convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664–675, 2015.

[21] Z. Xiong, M. K. Stiles, and J. Zhao, "Robust ECG signal classification for detection of atrial fibrillation using a novel neural network," in *Proceedings of the 2017 Computing in Cardiology (CinC)*, pp. 1–4, IEEE, 2017.

[22] T. Mahmud, S. A. Fattah, and M. Saquib, "Deeparrnet: An efficient deep cnn architecture for automatic arrhythmia detection and classification from denoised ECG beats," *IEEE Access*, vol. 8, pp. 104788–104800, 2020.

[23] H. Makimoto, M. Höckmann, T. Lin, D. Glöckner, S. Gerguri, L. Clasen, J. Schmidt, A. Assadi-Schmidt, A. Bejinariu, P. Müller, *et al.*, "Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction," *Scientific reports*, vol. 10, no. 1, pp. 1–9, 2020.

[24] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 535–541, 2006.

[25] O. Yildirim, M. Talo, E. J. Ciaccio, R. San Tan, and 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*, vol. 197, p. 105740, 2020.

[26] M. Baygin, T. Tuncer, S. Dogan, R.-S. Tan, and U. R. Acharya, "Automated arrhythmia detection with homeomorphically irreducible tree technique using more than 10,000 individual subject ECG records," *Information Sciences*, vol. 575, pp. 323–337, 2021.

[27] J.-H. Jang, T. Y. Kim, and D. Yoon, "Effectiveness of transfer learning for deep learning-based electrocardiogram analysis," *Healthcare Informatics Research*, vol. 27, no. 1, pp. 19–28, 2021.

[28] E. Debie, R. F. Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. Anavatti, M. Garratt, and H. A. Abbass, "Multimodal fusion for objective assessment of cognitive workload: a review," *IEEE transactions on cybernetics*, vol. 51, no. 3, pp. 1542–1555, 2019.

[29] G. K. Verma and U. S. Tiwary, "Multimodal fusion framework: A multiresolution approach for emotion classification and recognition from physiological signals," *NeuroImage*, vol. 102, pp. 162–172, 2014.

[30] M. Bansal and B. Gandhi, "Iot & big data in smart healthcare (ecg monitoring)," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 390–396, IEEE, 2019.

[31] K. C. Wee and M. S. M. Zahid, "Cloud computing for ecg analysis using mapreduce," in *2015 4th International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, pp. 115–120, IEEE, 2015.