DEAKIN UNIVERSITY

Impact of ECG dataset diversity on generalization of CNN model for detecting QRS complex

Citation:
Habib, Ahsan, Karmakar, Chandan and Yearwood, John Leighton 2019, Impact of ECG dataset diversity on generalization of CNN model for detecting QRS complex, *IEEE access*, vol. 7, pp. 93275-93285.

DOI: 10.1109/ACCESS.2019.2927726

© 2019, IEEE

Reproduced by Deakin University under the terms of the Creative Commons Attribution Licence

Downloaded from DRO:

http://hdl.handle.net/10536/DRO/DU:30128936

Impact of ECG dataset diversity on generalization of CNN model for detecting QRS complex

Citation:
Habib, Ahsan, Karmakar, Chandan and Yearwood, John Leighton 2019, Impact of ECG dataset diversity on generalization of CNN model for detecting QRS complex, *IEEE access*, vol. 7, pp. 93275-93285.

DOI: 10.1109/ACCESS.2019.2927726

© 2019, IEEE

Reproduced by Deakin University under the terms of the Creative Commons Attribution Licence

Downloaded from DRO:

http://hdl.handle.net/10536/DRO/DU:30128936

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 17, 2019, accepted July 2, 2019, date of publication July 10, 2019, date of current version July 29, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927726

Impact of ECG Dataset Diversity on Generalization of CNN Model for Detecting QRS Complex

AHSAN HABIB, (Graduate Student Member, IEEE),
CHANDAN KARMAKAR, (Member, IEEE),
AND JOHN YEARWOOD, (Member, IEEE)
School of Information Technology, Deakin University, Geelong, VIC 3225, Australia

Corresponding author: Chandan Karmakar (karmakar@deakin.edu.au)

ABSTRACT Detection of QRS complexes in electrocardiogram (ECG) signal is crucial for automated cardiac diagnosis. Automated QRS detection has been a research topic for over three decades and several of the traditional QRS detection methods show acceptable detection accuracy, however, the applicability of these methods beyond their study-specific databases was not explored. The non-stationary nature of ECG and signal variance of intra and inter-patient recordings impose significant challenges on single QRS detectors to achieve reasonable performance. In real life, a promising QRS detector may be expected to achieve acceptable accuracy over diverse ECG recordings and, thus, investigation of the model's generalization capability is crucial. This paper investigates the generalization capability of convolutional neural network (CNN) based-models from intra (subject wise leave-one-out and five-fold cross validation) and inter-database (training with single and multiple databases) points-of-view over three publicly available ECG databases, namely MIT-BIH Arrhythmia, INCART, and QT. Leave-one-out test accuracy reports 99.22%, 97.13%, and 96.25% for these databases accordingly and inter-database tests report more than 90% accuracy with the single exception of INCART. The performance variation reveals the fact that a CNN model's generalization capability does not increase simply by adding more training samples, rather the inclusion of samples from a diverse range of subjects is necessary for reasonable QRS detection accuracy.

INDEX TERMS Convolutional neural networks, deep learning, ECG, generalization, QRS complex, supervised learning, visual attention.

I. INTRODUCTION

Electrocardiogram (ECG) records the bio-electric response of heart's beating and characterizes a normal heart beat using a P wave, a QRS-complex and a T wave. The distinguishing shape of the QRS-complex forms the basis of ECG analysis [1], [2]. Detection of the QRS-complex may trigger the automated analysis of ECG characteristics (i.e., locate neighboring P and T waves, determination of R-R intervals, and heart rate), detection of cardiac anomalies [3], and classification of beats. ECG signal may also characterize individual subjects to form unique bio-metric signatures [4].

Over the last three decades, much research has been done on automated QRS detection. However, the challenges including the non-stationary nature of ECG, presence of different types of noise (e.g., baseline noise, power line interference, electrode contact noise, and motion artifacts) and signal variance at inter-patient, as well as, intra-patient recordings has kept the QRS detection as an active research area [1]. The traditional methodology of QRS detection is divided into a preprocessing stage and a decision stage. The preprocessing stage includes linear and non-linear filtering to suppress noise along with P and T waves and do feature extraction, whereas, the decision stage includes QRS detection and decision logic [1], [5]–[10]. Partial [9], [11] or all [5] records of publicly available databases (e.g., MIT/BIH arrhythmia, MIT-BIH in short) along with study specific databases were used in traditional approaches to report performance. Usage of a single database for performance evaluation leads to the fact that traditional approaches either lacked the generalization ability over unknown databases or such ability was not explored.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/

Convolutional neural networks (CNNs), a class of deep neural network, consist of convolution layers where each such layer automatically learns it's kernel coefficients during the training process through back-propagation. This constrained back-propagation enables CNN to achieve shift in-variance, robustness to distortions, reduction in free parameters and thus, requires proportionately smaller training data for certain levels of generalization performance with minimal preprocessing [12], [13]. In general, CNN inputs image data, however, one-dimensional (1-D) CNN is a variant of CNN which operates on time-series signals, like ECG signal. 1-D CNN is applied on ECG data for different tasks including beat classification [14]–[16], anomaly detection [17]–[20], QRS detection [21]–[26], sleep-wake classification [27] and bio-metric identification [4]. Most of the CNN-based QRS detection studies used single standard ECG databases (e.g., MIT-BIH database) for both training and testing [21], [26]. Although Xiang *et al.* [22] used two databases for analyzing generalization ability of a CNN model, it explored only one way generalization i.e., it trained using MIT-BIH database and tested on INCART database, but not vice versa. In addition, the effect of multi-database training on generalization ability of CNN in QRS detection is yet to be explored. Therefore, further research is warranted for analyzing the generalization ability of CNN in QRS detection.

The main contribution of this study is to investigate the impact of ECG dataset diversity on generalization of a CNN model, both intra and inter-database testing approaches using three publicly available datasets were applied. Intra-database validation is used to analyze subject-wise generalization ability of the model within individual database. Whereas, inter-database validation reflects the generalization ability of the model beyond the training database(s) where validation is performed using unknown database(s). A new CNN model is proposed to carry out this investigation following an existing best performing CNN model from the literature [22] for QRS detection.

The rest of this paper is structured as follows. Section II presents the methodology adopted for QRS detection generalization problem. Section III describes ECG databases. Obtained results are presented in section IV and analyzed in section V. Finally, section VI concludes the paper.

II. METHODOLOGY

A. SEGMENTATION AND DIFFERENTIAL SIGNAL ESTIMATION

An ECG signal segmentation strategy aims to capture the QRS morphology. For heart beat detectors, like this study, a segment is to be defined along with a proper sliding measurement so that it can maximize the likelihood of encompassing the QRS morphology of almost all of the beats leaving fewer beats with, at least, partial QRS structure. According to ANSI/AAMI EC38 and EC57 standards [28], while localizing heartbeats, an estimated location is deemed accurate if it is no further than 150 ms from the corresponding annotated location. This means, the current sample point can be represented as an R peak if the annotated peak remains in the range of 150 ms before or after the current detection point. In [21], a similar philosophy was followed in segmentation. The ECG databases in this study are sampled at 360 Hz and 150 ms is equivalent to 54 samples which when considered on either side of a detection point, forms a total of $108+1 = 109$ samples. The group of 109 samples forms a detection window and is also called an inner segment, in this study, which can be represented as $54 + 1 + 54 = 109$ samples. A segment envelope is now formed around this inner segment by appending 150 ms equivalent samples at either end of the inner segment having an orientation of $54+54+1+54+54 = 217$ samples. The terms *envelope* or *segment* will be used to refer to the full segment of 217 samples and the *inner segment* or *detection segment* will refer to the 109 sample segment. The segment is then shifted keeping no overlapping of inner segments. Figure 1 shows segmentation, segment shift and the network architecture. There are two reasons behind keeping no overlap of inner-segments. The first reason is obviously not to detect

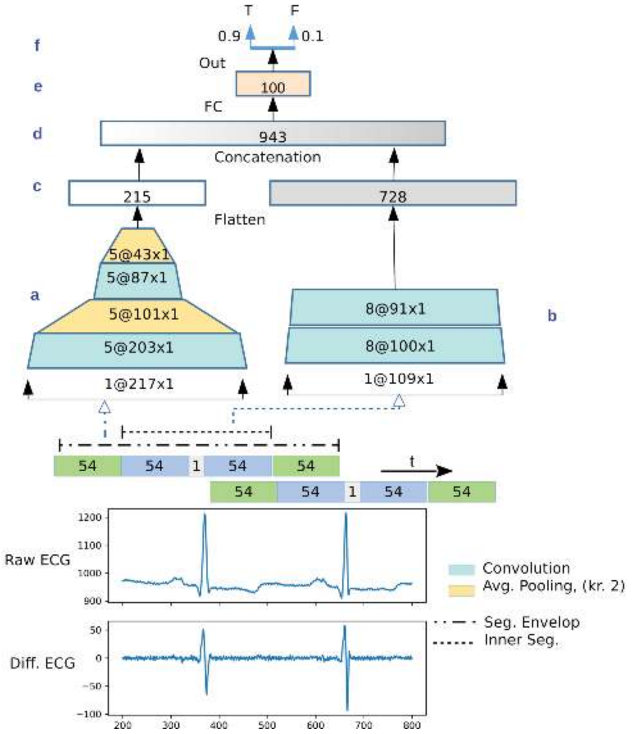

FIGURE 1. The proposed new model architecture of two-level attention-based CNN. In the figure, (a) is the first level CNN which consists of two sets of alternate convolution (cyan) and average-pooling (yellow) layers, (b) is second level CNN consisting of two consecutive convolution layers with no pooling layer, (c) represents flattened CNN features (number of channels * features per channel) of each level CNN, (d) is concatenation of two levels flattened features which is fed to fully-connected layer (100 neurons) in (e) followed by the output layer in (f) which consists of two neurons. The bottom part of the figure shows the formation of a segment from raw and differentiated ECG signals. The dash-n-dotted line represents a whole segment, known as signal-envelope, and dotted-line represents inner segment with the middle point as the R-peak detection point. A signal-envelope moves forward in time with $54 * 2 = 108$ samples overlap.

the same R-peak again in the next shifted segment and the other reason is that as the R-R interval, on average, is greater than 0.6 seconds, next QRS is likely not to occur before at least 0.6 seconds of the previous occurrence, so the detection segment will not miss that event due to this non-overlapping decision.

The experiment feeds differentiated ECG signal segment-by-segment into the CNN model. By differentiating ECG signals, a QRS-complex becomes more prominent, appearing as a high slope and the resultant signal spans around the zero axis. Differentiation of raw ECG signal is considered as a minimum level of preprocessing [22]. In this study, the raw ECG signal was segmented progressively and just before feeding into the network, the current segment was differentiated. This just-in-time differentiation applied to a signal segment facilitates immediate feature extraction by the proposed approach as opposed to many other approaches where mandatory preprocessing steps were required to apply to the overall signal before any kind of feature extraction. In this study, equation 1 was used to differentiate a raw ECG signal segment

$$y_1(n) = x(n) - x(n-1) \tag{1}$$

where a difference signal yielded by subtracting previous sample amplitude from that of the current sample.

B. ATTENTION BASED HIERARCHICAL AND MULTI-SCALE CNN

To analyze the generalization capability, a suitable model is needed. There are two choices, either select an existing established model or create a new model with comparable performance.

A shallow CNN architecture [21] (single convolution layer) with multi modal physiological inputs (i.e., arterial blood pressure along with ECG Signal) shows good performance where more than two stages of preprocessing were applied to the input signal. In our study, CNN is expected to perform well with a minimum level of preprocessing and a single physiological input (i.e. ECG Signal), therefore, considering these facts, that model was not selected. Another CNN-based QRS detection study [26] takes a single ECG signal as input and reported 99.81% sensitivity with MIT-BIH arrhythmia database by using a shallow network structure of two convolution and one pooling layers. In that study, each sample point was considered as a detection point which was described by a sample of 145 neighboring points (considering 360Hz sampling frequency of MIT-MIH arrhythmia database) and due to a single point shifting, several positive predictions for a single QRS detection was eventually generated which was handled by applying an optimized clustering by taking all the classification outcomes at the end of each subject recordings. Due to the existence of clustering constraint with all the accumulated classification decisions, this model was considered not suitable in this study where the main focus is to analyze the strength of a CNN only without mixing any other machine learning methods. A *Two-level* attention-based CNN model [22] reported high accuracy and positive prediction rate (PPR), and for training the model, 400 representative QRS complexes along with associated non-QRS segments were selected. Although this model used minimal preprocessing, it could not be reproduced due to the lack of information regarding the selection of 400 representative beats. At this point, although the main focus of this study is to investigate the generalization capability of a CNN model and not to find the best CNN model for QRS detection from ECG signals, no existing CNN model from literature, for the scope of this study, could be selected and the only option left was to create a new CNN architecture which performs at least, as good as the above CNN models.

Towards finding a new model, the attention-based two-level CNN model [22] was the inspiration in this study due to the fact that this shallow CNN architecture well performed among other two CNN models discussed above for QRS detection and in addition, model's performance (trained on MIT-BIH arrhythmia database) was validated against a second database (i.e. INCART database) which was unknown to the model. There were two CNN sub-networks (a.k.a. levels) which took ECG segments of two different scales as inputs to extract two streams of features. For the first level CNN (known as object-level CNN), an average operation of five ECG samples was performed, followed by a difference operation of each averaged outcome to form a segment of average-difference signals as input which then went through two sets of convolution-pooling layers to extract coarse-grained features. The input segment for second layer CNN (known as part-level CNN) was formed by simply taking difference operation between successive samples which then went through a single convolution-pooling layers to extract fine-grained features. Two streams of features were then combined and then fed into the classifier (i.e., fully connected layer) for classification. One of the confusions regarding that model was why five samples were taken for average operation instead of any other number of samples. While designing a model for this study following the above two-level attention-based model, alternative approaches were investigated for the object-level CNN to address the above mentioned confusion. To keep things simple, in this experiment, the idea of averaging the samples was discarded, and same difference signal was used for both the object and part-level CNNs but with different lengths. Using trial-and-error method the following model decisions were reached:

- A segment with double number of samples for the object-level CNN input than the part-level CNN was fixed,
- The object-level CNN contained two sets of convolution-pooling layers and was optimized to have larger receptive fields than the part-level CNN,
- The part-level CNN contained no pooling layer, instead, it had two successive convolution layers.

TABLE 1. Data flow through different layers and model parameters of proposed multi-level CNN of ECG signals sampled at 360 Hz. Here, Conv: Convolution layer, Pool: Average pooling, BN: Batch normalization, FC: Fully-connected layer.

CNN Level	Layer	In	#Channels	Kernel	Out
Object (seg. env.)	Input	217	1	-	-
	Conv	217	5	(15, 1)	203
	Pool	203	-	/2	101
	BN	-	5	-	-
	Conv	101	5	(15, 1)	87
	Pool	87	-	/2	43
	BN	-	5	-	-
	Flatten	-	-	-	215
Detailed (inner seg.)	Input	109	1	-	-
	Conv	109	8	(10, 1)	100
	Conv	100	8	(10, 1)	91
	Flatten	-	-	-	728
	Concat	-	-	-	943
	FC	-	-	-	100
	BN	100	1	-	-
	Dropout	0.5	-	-	-
	Out	-	-	-	2
Learning rate: 0.001, Loss fn: Cross Entropy, Optimizer: Adam.					

In the two-level CNNs, one CNN level extracts coarse grained object-level abstract view of the signal and the other level CNN extracts fine grained detailed features from the inner segment of the concerned segment envelope. These hierarchical views are then combined and sent to the fully connected (FC) layer followed by a two-neuron output layer for decision making - a positive outcome if a QRS event is present in the inner-segment or a negative outcome otherwise. This scheme is illustrated in figure 1 and network details with changes in the input-output volume are summarized in table 1.

In the literature, multi-scale was found to achieve, among other methods, using pooling operations, dilated convolution [29], manual scaling operations [22], [30], and different interpolations [31]. In this study, multi-scale effect might be observed in several places in several ways, including (i) the case of the part-level CNN which utilizes more convolution filters than the object-level CNN, thus producing increased number of feature-maps, (ii) the case of object-level CNN which utilizes two pooling layers but the part-level CNN has no pooling layer, (iii) the case of the input size of the object-level CNN that is double than the part-level CNN, and finally (iv) the case of the object-level CNN that has larger receptive fields than the part-level CNN. The CNN uses a learning rate of 0.001 throughout all epochs, a cross-entropy loss function, and an Adam optimizer. The advantage of using cross-entropy is that the natural log function takes into account the closeness of a prediction and is a more granular way to compute error. The Adam optimization algorithm [32] is used instead of the classical stochastic gradient descent procedure as it is efficient and performs better with minimal tuning of input parameters. To make sure network weights are in a reasonable range before training starts, they are initialized with Xavier uniform initialization [33] to get better training performance. The ECG segment envelope and inner detection segment consists of 217 and 109 samples respectively which places a limit on the number of CNN layers to be used in the architecture.

1) SEGMENT ENVELOPE: OBJECT LEVEL ABSTRACT VIEW

The two-level CNN consists of two CNN sub-networks. The first sub-network inputs a signal envelope and focuses on object level features whereas the other sub-network inputs an inner-segment of a signal envelope and focuses on detailed features. The object-level sub-network consists of two identical layer sets where each layer set consists of a convolution layer with five filters (a.k.a. channels) followed by a non-linear activation (ReLU) and sub-sampling (a.k.a. pooling) layer. In general, three main layers are used to build the CNN architecture: a convolutional layer, a pooling layer, and a fully-connected layer. Each neuron in a convolution layer does not connect to every neuron in the previous layer, rather each neuron is connected to only a local region of the input volume. The spatial extent of this connectivity is a hyper-parameter called the receptive field of the neuron which equivalently is the filter or kernel size. Each convolution layer in the object-level CNN uses five filters of fifteen samples long receptive fields where each filter convolves through the segment envelope and produces a feature-map, yielding five convoluted feature maps in total. Changing the receptor field size or the number of filters does not always yield increased performance. Filtering with greater or smaller kernels has corresponding side effects. By increasing the kernel size of a filter, a neuron basically becomes responsible for summarizing a larger receptive field of the previous layer yielding a smaller convoluted feature map which in the case of time-varying signal data may fail to reveal interesting patterns. On the other-hand, decreasing kernel size of a filter yields a larger convoluted feature map which may possibly capture unnecessary information of the signal representing noise. The number of convolution layers per level for two-level sub-networks, the number of filters per convolution layer and each filter's kernel size are optimized based on the MIT-BIH database using K-fold validation approach of $K = 5$. A filter convolves with the receptive field in the previous layer input and produces single scalar output and then slides to the next adjacent input data with a certain overlap. This sliding operation is known as a stride and in this study a stride is one. The filter slides one sample to the right to produce a single scalar output and repeating this process eventually yields a convoluted feature map. Each hidden unit in a feature map shares parameters (weights and biases) with all other hidden units in that feature map to compute their pre-activations reducing the number of parameters dramatically. Sharing parameters also means that hidden units of a feature map are extracting the same feature from the previous layer. It might be tempting to extract lots of features by using lots of filters, however, computing resources restrict such an approach and there is a performance threshold beyond which there is no further improvement observed. As in [34], the computation of hidden layer activation can be summarized as

$$y_j = g_j \; tanh(\sum_i k_{ij} * x_i) \tag{2}$$

where x_i is the i^{th} channel of input, k_{ij} is the convolution kernel, g_j is a learned scaling factor, and y_j is the hidden layer. Pre-activations are computed using the convolution of each of the i^{th} input channel to the j^{th} feature map and then summing across that input channel. Then a non-linear *tanh* function is applied among other alternatives such as the sigmoid, and the rectified linear unit (ReLU). Optionally, a bias could also be introduced as b_j to be shared across all the feature maps.

It is common to periodically insert a pooling layer in-between successive convolution layers which progressively reduces the spatial size of the representation to reduce the number of parameters and the computation in the network. In addition, the pooling layer helps to control over-fitting. Each convolution layer of the object-level CNN follows a sub-sampling layer with a (2, 1) filter which produces output size of half of the input. There are mainly two kinds of sub-sampling operations - max and average sub-sampling. Max sub-sampling outputs the maximum value of the clustered neurons, and average or mean sub-sampling outputs the average value of the clustered neurons. In general, max sub-sampling picks prominent features and mean sub-sampling moves combined affects forward. In this work, mean-pooling was used as it performed slightly better than max-pooling, even though max-pooling is commonly used on image input data. This network extracts coarse and spatially-varying sets of features from the signal envelope. Figure 1 represents input-output dimensions at every step of the CNN architecture.

2) INNER SEGMENT: FINE-GRAINED DETAILED VIEW

The second sub-network of the CNN architecture consists of two convolution layers only with no pooling layer and inputs the inner detection segment of a segment envelope. In order to extract fine grained detailed features, eight convolution filters were used instead of five that were used in the object-level sub-network. If an annotated peak exists within this region, then the model should output a positive decision, that is why this region is called the detection segment in this study. It is obvious to notice that a QRS may appear at any point within this region, not necessarily at the middle detection point. Sometime it may happen that the R-peak almost touches the detection boundary and the detection segment contains a partial QRS structure. So, the features extracted from this inner segment may contain QRS morphological features completely or partially. This part-based detection philosophy already exists in the computer vision literature [35], [36] and a similar philosophy was explored here as well. In those studies, an image was segmented into several regions (a.k.a., bottom-up attention) followed by a top-down filtering operation (a.k.a., top-down attention) to filter out regions containing no detectable object and then extract part-level discriminating features from these regions to make fine-grained detection decisions. In this study, a simple one-dimensional ECG signal is used and the inner-segment is considered as a single proposed region from which part-based discriminating features of QRS are extracted which takes part in the detection decision in combination with object level features extracted from the segment envelope. Output features from both CNN sub-networks are concatenated and sent as input to the fully connected (FC) layer and finally to the output layer of two neurons which declares the presence or absence of QRS.

C. TRAINING AND TESTING APPROACHES

In this study, two different testing strategies are used - intra database and inter database testing which are described in the following subsections. The segmentation process of ECG data produced more negative labeled segments, almost double, than positive labeled segments per ECG records where a positive labeled segment contains a QRS event, however, negative labeled segments do not. The increase of class imbalance generally has an adverse effect on a classifier's test performance [37] and this, in general, is compensated by either under-sampling the majority class or supplement the under-sampling operation with over-sampling of the minority class [38]. The former approach among the two, under-sampling the negative labeled segments, is adopted to maintain class balance for training purposes by removing randomly chosen negative labeled segments per ECG recording. Note that during testing, all segments from recordings are sent to the network as a continuous stream of segments without any filtering.

1) INTRA-DATABASE TESTING

Intra-database testing was carried out in order to validate generalization ability within individual databases. This type of testing was performed using *leave-one-out* cross testing and *k-fold* cross testing with $k = 5$ on each database. Leave-one-subject-out testing uses all the segments from $n - 1$ recordings per database for training and then validates on the segments from the remaining n-th recording. Five-fold validation is carried out subject-wise. In subject-wise five-fold validation, recordings from each database are divided into five folds and then testing is carried out over each fold sequentially, using the remaining $k - 1 = 4$ folds for training. For example, if a data set contains 46 recordings, then 5-fold cross validation will generate five training and testing sets composed of [9, 9, 9, 9, 10] and [37, 37, 37, 37, 36] subjects respectively.

2) INTER-DATABASE TESTING

Inter-database testing was used to analyze the generalization ability beyond the training database(s) by using one or more unknown testing database(s) and was carried out in two phases. The first phase considers one database for training and the other two for testing. All the recordings from a database were segmented, a balance of class segments was achieved by removing randomly selected negative labeled segments and the model was trained. For testing, all

TABLE 2. Used performance metrics.

$Sen(\%) = \frac{TP}{TP+FN} * 100$	where, TP = no. of true positives
$Spe(\%) = \frac{TN}{TN+FP} * 100$	TN = true negatives
$PPR(\%) = \frac{TP}{TP+FP} * 100$	FN = false negatives
$Acc(\%) = \frac{TP+TN}{TP+TN+FP+FN} * 100$	FP = false positives

recordings from each target database were segmented and fed into the network for testing without any removal of segments. In the second phase of inter-database testing, a combination of two databases were used for training and then validated with the single unknown database. The training dataset was prepared similarly by segmenting all the records from two source databases, a class balance is maintained by removing randomly selected negative samples from each record per database, subjects from both databases are mixed together by pulling the records from both databases alternatively and then sent into the network for training. All the segments of the target database are then used for testing with no exclusion of segments.

D. PERFORMANCE METRICS

Metrics used to measure the performance of both intra and inter-database testing scenarios are sensitivity (Sen), specificity (Spe), positive predictive rate (PPR), and accuracy (Acc). These metrics are formulated as Table 2.

E. MODEL IMPLEMENTATION

To access ECG data files, segment and differentiate the records, implement the CNN model, perform training and validation of the model, *PyTorch* [39] was used in this study which is a Python-based deep-learning library. It provides an easy and intuitive way to define and process a dataset, design and tweak the neural network and above all, it is relatively easy to code and debug.

III. ECG DATA

Three ECG databases from the Physionet data bank [40] are used in this study-the MIT-BIH Arrhythmia database (MIT-BIH), the QT database, and the St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database (INCART). Each database has unique characteristics and is sampled at different frequencies. For instance, the MIT-BIH, INCART and QT comes with ECG recordings of 360, 257, and 250 Hz sampling frequencies respectively and different ADC gains and formats. In this study, experiments are carried out on these databases at 360 Hz sampling frequency to ensure all samples share common recording characteristics. In order to do that, recordings from the later two databases, INCART and QT, were re-sampled at 360 Hz. All beats annotated with the American Heart Association (AHA) standard set of annotation codes (along with further sub-division of Physionet) were considered except the paced (P) and fusion (f) of paced and normal beats. The databases are summarized in Table 3.

TABLE 3. Characteristics of Physionet databases used in this study.

DB Name:	MIT-BIH	INCART	QT
Source Hz:	360	257	250
Target1 Hz:	360	360	360
# Records:	48	75	105
Used # Rec:	48	75	82
# Leads:	2	12	2
Used Lead:	Channel-1 (MLII or V5)	Channel-1	Channel-1
Rec Len(minute):	30	30	15
# Beats:	101,509	175,461	84,386
# Segments:	275,215	444,611	239,381
-ve segments:	173,706	269,150	154,995
+ve segments:	101,509	175,461	84,386

The MIT-BIH contains 48 half-hour excerpts of two-channel ambulatory recordings from 47 subjects and each recording is sampled at 360 Hz with 11-bit resolution over a 10 mV range. Among the two channels, common is the modified-lead II (MLII) except two recordings, record 102 and 104, and the other lead is mainly lead V1, sometimes V2, V4 or V5. This study uses MLII lead signal whenever available, otherwise the first lead among the available leads is used. This contrasts several studies where recordings of only MLII leads were used and above mentioned two records were excluded [22]. In this study, beats from records 102, 104, 107, 217 were taken partly excluding paced and fusion beats. The INCART consists of 75 annotated recordings extracted from 32 Holter records where each record is 30 minutes long and contains 12 standard leads, each sampled at 257 Hz. In this study, recording from the first lead (lead I) was used for all the INCART records. The third database, QT, consists of 105 fifteen-minute excerpts of two-channel ECG Holter recordings, each sampled at 250 Hz, and include a broad variety of QRS and ST-T morphology. Recording from the first channel of the two was used in this study. Among 105 recordings, 23 records whose names range from sel-30 to sel-52 in sequence were not provided with annotation files (.atr file) with them and this is why these recordings were excluded from this study leaving only 82 recordings to work with. An annotation file contains sets of labels which point to specific locations in the recording and describe features at those points, indicating signal sample location and type. Physionet tool *xform* was used to re-sample signals to target frequency.

IV. RESULTS

A. INTRA-DATABASE TESTING

1) LEAVE-ONE-OUT (LOO) TESTING

The accuracy of leave-one-out testing on three databases are represented in the bar charts in Figure 2-A and the Table 4. The MIT-BIH shows the highest accuracy, followed by a decrease in INCART and the lowest in QT. In addition, the standard deviation is minimum in MIT-BIH (less than 1%) and almost equal (2.7%) in INCART and QT.

2) K-FOLD TESTING

The accuracy of five-fold testing performance of three databases are shown in the Figure 2-B and Table 4. Similar to

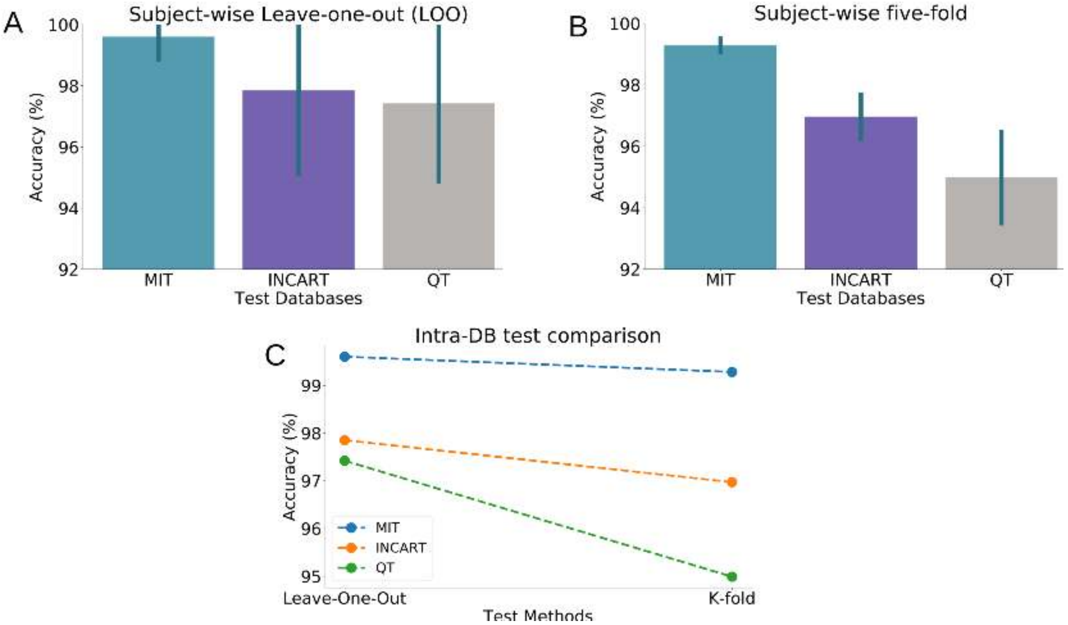

FIGURE 2. Intra-database subject-wise test accuracy of (A) Leave-one-out (LOO), (B) K-fold (five-fold) and (C) comparison between these two test methods. Databases are MIT-BIH Arrhythmia, INCART, and QT.

TABLE 4. Intra-database test (e.g., leave-one-out (LOO) & K-fold) metrics for QRS detection. Databases are MIT-BIH Arrhythmia, INCART, and QT.

Database	Test	Sen (%) mean std	Spe (%) mean std	PPR (%) mean std	Acc (%) mean std
MIT-BIH	LOO	99.22 ±1.84	99.75 ±0.45	99.38 ±1.28	99.60 ±0.81
	K-fold	98.68 ±0.69	99.55 ±0.17	97.93 ±2.52	99.28 ±0.29
INCART	LOO	97.13 ±4.02	98.36 ±1.81	97.70 ±2.38	97.85 ±2.79
	K-fold	94.90 ±1.98	98.07 ±0.39	96.54 ±0.77	96.97 ±0.79
QT	LOO	96.25 ±3.90	97.91 ±2.73	96.81±2.95	97.42 ±2.61
	K-fold	92.91 ±2.26	96.04 ±1.33	91.31 ±5.26	94.99 ±1.56

TABLE 5. Inter-database test (e.g., training with single-db & combined-db) metrics for QRS detection. Databases are MIT-BIH Arrhythmia, INCART, and QT.

Train DB	Test DB	Sen (%)	Spe (%)	PPR (%)	Acc (%)
MIT-BIH	INCART	85.09	95.32	92.49	91.65
	QT	92.79	95.89	91.32	95.04
INCART	MIT-BIH	97.61	97.43	91.93	97.49
	QT	92.95	94.95	89.28	94.39
QT	MIT-BIH	94.94	97.38	92.37	96.6
	INCART	77.89	90.76	86.07	86.3
INCART + QT	MIT-BIH	97.8	97.38	92.1	97.5
MIT-BIH + QT	INCART	85.26	93.02	90.33	90.54
MIT-BIH + INCART	QT	93.45	94.8	89.68	94.48

leave-one-out testing, the MIT-BIH shows the highest accuracy, followed by a decrease in INCART and the lowest in QT. However, compared to leave-one-out, accuracy decreased in *K-fold* validation where MIT-BIH variation was negligible compared to INCART (1%) and QT (2.5%) (Table 4). The variances in *K-fold* are smaller than leave-one-out across all databases. This could be due to the fact that deviation in K-fold represents the variation across folds rather than individual subjects.

The comparison of the accuracy between two intra-database testing methods namely LOO and K-fold is shown in Figure 2-C. It is obvious that the k-fold testing shows worse performance than leave-one-out testing for all databases. This decrease in performance can be attributed to the lower number of training samples. Interestingly, the fall in performance (the slope) is not the same across all databases and it is highest for QT.

B. INTER-DATABASE TESTING

1) TRAIN ON ONE-DATABASE, TEST ON OTHER-TWO

The accuracies of the first phase of inter-database testing are shown in the Figure 3-A and Table 5 where samples of a single database were used for training and the model was tested on samples of other two databases. The CNN model trained using the MIT-BIH database shows better accuracy for both the QT and INCART database (Figure 3-A) with the QT database slightly higher than the INCART database. On the other hand, the CNN model trained using the INCART database shows best accuracy for the MIT-BIH database and slightly lower for the QT. The QT database was nearly equally generalized by both the MIT-BIH and the INCART databases. Finally, although, the QT trained model generalized the INCART database worst, it generalized the MIT-BIH database nearly equally as the INCART did.

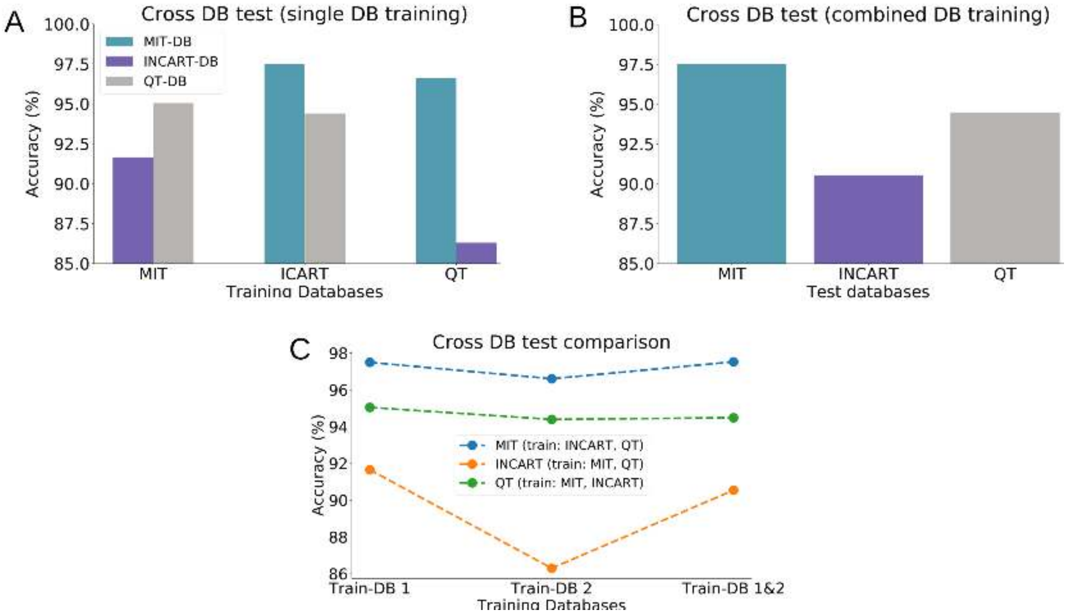

FIGURE 3. Cross-database testing accuracy of ECG databases using (A) single database training (CNN model trained with single database used to test other two databases separately), (B) two database joint-training and (C) comparison between these two. For example, in (A), first bar-pair represents testing databases (INCART, QT) for which MIT database was used for training, so the label 'MIT' stands for.

2) TRAIN ON TWO-DATABASES, TEST ON THE REMAINING

The accuracies of the second phase of inter-database testing are shown in Figure 3-B and Table 5 where a pair of databases were used to train the model and validated on the remaining one. The testing performance of the MIT-BIH, when trained with the other two databases (e.g., INCART and QT), is the highest, closely followed by the QT and the lowest INCART. The comparisons of accuracy of two inter-database validation processes are summarized in Figure 3-C. The accuracy consists of three different colored-lines which represent three different test databases where each colored-line consists of three points. First two points in a line represent the validation performance of a database when trained with each of the remaining databases individually, whereas, the third point indicates validation performance when model is trained with the other two databases combined. The figure shows that the model's performance on a validation database, when trained with other two databases combined, does not go beyond the best performance on that validation database when trained individually.

V. DISCUSSION

In this study, a two-level attention-based new CNN model was proposed and the generalization capability of the model was explored using intra and inter-database testing over three publicly available ECG databases namely MIT-BIH, INCART and QT. Intra-database testing includes subject-wise leave-one-out and K-fold testing, whereas, inter-database testing considers single and multi-database training. Intra-database test reveals that the CNN model better generalizes subjects in a database which has lesser subject-level variance. On the other hand, the inter-database test shows that increasing the volume of training samples even from multiple databases does not increase accuracy beyond the best accuracy which is achieved using training databases individually.

TABLE 6. Distribution of the 105 records according to the original database.

Database		Records
MIT-BIH	Arrhythmia	15
MIT-BIH	ST DB	6
MIT-BIH	Sup. Vent.	13
MIT-BIH	Long Term	4
MIT-BIH	NSR DB	10
ESC	STT	33
Sudden	Death	24

Intra-database testing accuracy (both LOO and K-fold testing) show better generalization capability of MIT-BIH database (Figure 2). This indicates that the model better generalizes over the unseen subjects of MIT-BIH database than either INCART or QT. This may be due to lower inter-subject variation of MIT-BIH database compared to other two databases. The result also shows that the accuracy of the model decreases consistently in K-fold testing across all databases compared to LOO testing which may be attributed to the reduced number of training samples from LOO to K-fold test. However, this decrease in accuracy observed in MIT-BIH is negligible (0.3%) compared to INCART (0.9%) and QT (2.4%). This varied decreasing rate of accuracy reveals the fact that even though the reduced training samples have an impact on generalization, the characteristics of

TABLE 7. Comparison of generalization performance (only sensitivity is shown) of several QRS detection methods (i.e. both traditional and CNN based) on different databases (DBs), including MIT-BIH Arrhythmia (MIT-BIH), INCART, and QT. Intra-database validation methods include Leave-one-subject-out (LOO) and K-Fold, whereas inter-database validation refers to a model validated with one or more test databases which was trained using a single or combined databases.

Method	Train DB			Test DB Sen(%)		
	MIT-BIH	INCART	QT	MIT-BIH	INCART	QT
Hamilton et al. [7]	x			99.69		
Sarlija et al. [26]	x			99.81		
Chandra et al. [21]	x			99.94		
Xiang et al. [22]	x			99.77	99.86	
This work	x			99.22 (LOO) 98.68 (k-fold)	85.09	92.79
		x		97.61	97.13 (LOO) 94.90 (k-fold)	92.95
			x	94.94	77.89	96.25 (LOO) 92.91 (k-fold)
	x	x				93.45
		x	x	97.8		
	x		x		85.26	

the subjects held for testing might have a greater impact. Databases affected by the holding of subjects likely to contain recordings from a diverse range of subjects. Thus, holding one-fold subjects, containing unique signal characteristic, from training may restrain the CNN model to generalize well over this fold. The MIT-BIH's negligible decrease in accuracy further supports the finding that its subject level variance is smaller than the INCART and the QT. The maximum decrease in accuracy in the QT may be attributed to the diverse level of subjects it contains. Further investigation to claim the presence of diversified subjects in QT reveals that the QT database is composed of different other databases (Table 6) of varied subject characteristics [41]. This finding affirms the claim that the subject-level variance being a major cause of poor generalization in intra-database testing. Thus, exposure of the CNN model to diverse training samples is important than more samples of similar type.

Inter-database testing reveals that the MIT-BIH database is best generalized by either of the training databases (more than 96.5%, Figure 3, Table 5). This may follow similar justification as the intra-database testing that the MIT-BIH recordings likely to have lesser subject level variance which INCART or QT database-based trained model finds easier to generalize. The QT database is the second best database that has been generalized nearly equally by the MIT-BIH and INCART (95% and 94.4%). However, this accuracy is lower than MIT-BIH generalization accuracy. The reason for comparatively poor generalization of QT database probably due to the fact that it contains recordings from diverse range of subjects for which the model, trained with either MIT-BIH or INCART, had difficulty in QRS detection. The existence of such diversified subjects in QT database can be understood from its composition as shown in Table 6. Although the INCART database has greater intra-database accuracy than QT, interestingly, it had the poorest generalization by QT (86.3%), as well as MIT-BIH (91.6%). There might be subject level unique signal characteristic of INCART that hinders its generalization. Moreover, several recordings of the INCART database (Figure 4 D-F) show much noise which is likely to have a major impact on this poor generalization. The MIT-BIH database is comparatively less noisy than the other two databases (Figure 4 A-C) and this characteristic, in addition to having lesser subject level variance, may have additional influence to its generalization by noisy databases. It looks like the ECG database with noisy recordings are able to better generalize databases with comparatively less noisy recordings, although, the presence of lesser varied subjects, as well as, similar recordings (e.g., presence of 15 MIT-BIH records in QT) may show biased performance.

In another scenario where the model was trained using samples from two databases (INCART and QT), the test accuracy of MIT-BIH (97.50%, in Table 5) did not increase compared to single database training INCART (97.49%) or QT (96.60%). Moreover, the combined database training yields test accuracy for QT (94.48%) and INCART (90.54%) slightly below the best accuracy (95.04% and 91.65% respectively) achieved using the single database (MIT-BIH for both cases) training 5). Therefore, combining samples of multiple databases does not better generalize the model than that is achieved using the model trained using one database (Figure 3-C). This indicates that inclusion of a second database in training is adding complementary knowledge rather than supplementary (for this set of databases) and hence not aiding in improving accuracy. In addition, the presence of different types of noise (e.g., baseline noise, power line interference, electrode contact noise, and motion artifacts) in the test database can make it difficult to achieve higher accuracy using a model trained with less noisy databases. In particular, a closer look into comparatively noisy signal patterns of INCART recording number 12 and 53 (Figure 4-E, F) probably justifies this claim.

The performance of a QRS detector should consider the problems of noisy or pathological signals [1]. The scenario of inter-database testing stresses the CNN model to explore some of these problems. The inter-database testing accuracy of both single database and combined database training shows that the model generalizes to unknown databases with more than 90% accuracy (except INCART that was poorly generalized in both categories). The CNN model generalization studies over multiple databases are scarce in the QRS

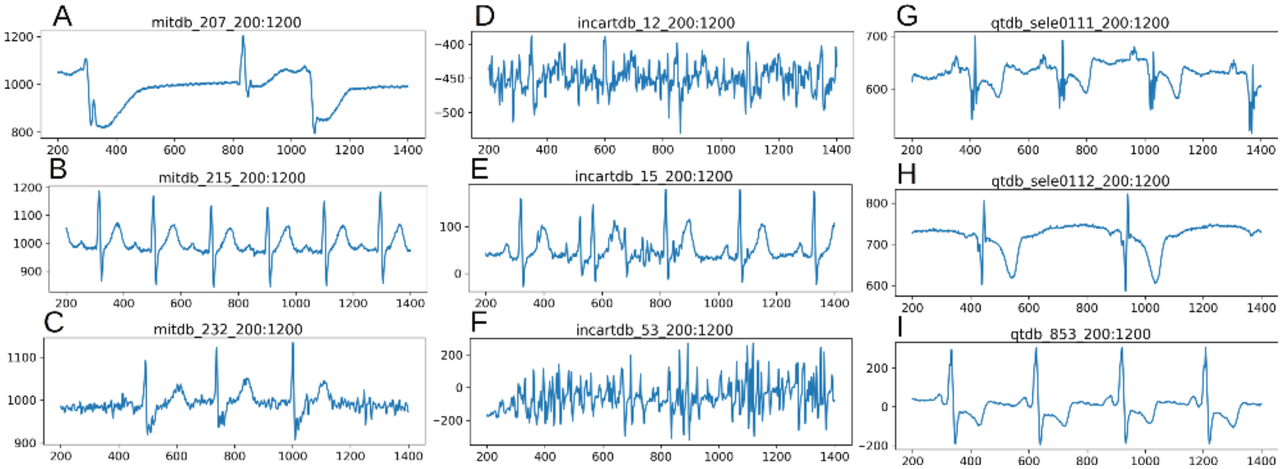

FIGURE 4. Comparatively noisy ECG recording excerpts per database, (A-C) MIT-BIH, (D-F) INCART, and (G-I) QT. Each signal excerpt consists of samples ranging from 200 to 1200. Recording number of each signal is mentioned at the top of corresponding excerpt.

detection context. Table 7 presents the sensitivity of a single traditional and three CNN based QRS detection methods along with results of this current study. In a QRS detection study [22], the test sensitivity on the INCART database was reported as 99.86% (Table 7) on a CNN model that was trained using a subset of the MIT-BIH database, however, the subset selection and training process was not explicit which makes the result difficult to compare. Moreover, the accuracy of the opposite scenario was not shown (training on INCART and test on MIT-BIH). The accuracies in this study were achieved with minimal level of preprocessing by only differentiating the ECG signal. Increased performance may have been observed if common preprocessing steps (e.g., baseline correction, removal of power-line & high-frequency noise etc.) were applied to the input signal. The reason for not using extensive preprocessing is simply because this study intended to investigate the learning ability of CNN with minimal preprocessing which is one of the strengths of CNN [12], [13]. For the combined database training, test performance does not increase by adding more samples of similar type, rather, a balanced ECG signal collection including diverse subject types for training may help achieve reasonable performance across different test databases.

VI. CONCLUSION AND FUTURE WORK

In this study, a new CNN architecture was proposed for QRS detection whose intra and inter-database based generalization capabilities were tested on three publicly available ECG databases to explore its dependency on dataset characteristics. The accuracy of both test categories show reasonable generalization accuracy (more than 90%) with single exception of one database (i.e., INCART) which falls behind in the inter-database test category. In addition, the accuracies were achieved with minimal preprocessing by taking only a difference of the raw ECG signal. The study reveals the fact that the CNN model's generalization performance does not increase by simply adding more training ECG samples of similar subjects but a diverse range of subjects should be included. This also shows that the high accuracy obtained using the intra-database testing approach does not reflect the true generalization capability of a CNN model. In the future, we aim to explore different CNN architectures (including but not limited to multi-dilated convolution, and cross-layer feature aggregation) to find the best CNN model for QRS detection and further investigate the generalization capability against minimum required training sample size and diverse subjects with justified level of preprocessing.

REFERENCES

[1] B.-U. Kohler, C. Hennig, and R. Orglmeister, "The principles of software QRS detection," *IEEE Eng. Med. Biol. Mag.*, vol. 21, no. 1, pp. 42–57, Jan./Feb. 2002.

[2] F. Bouaziz, D. Boutana, and M. Benidir, "Multiresolution wavelet-based QRS complex detection algorithm suited to several abnormal morphologies," *IET Signal Process.*, vol. 8, no. 7, pp. 774–782, Sep. 2014.

[3] H. L. Chan, W. S. Chou, S. W. Chen, S. C. Fang, C. S. Liou, and Y. S. Hwang, "Continuous and online analysis of heart rate variability," *J. Med. Eng. Technol.*, vol. 29, no. 5, pp. 227–234, 2005.

[4] Q. Zhang, D. Zhou, and X. Zeng, "HeartID: A multiresolution convolutional neural network for ecg-based biometric human identification in smart health applications," *IEEE Access*, vol. 5, pp. 11805–11816, 2017.

[5] C. Li, C. Zheng, and C. Tai, "Detection of ECG characteristic points using wavelet transforms," *IEEE Trans. Biomed. Eng.*, vol. 42, no. 1, pp. 21–28, Jan. 1995.

[6] P. S. Hamilton and W. J. Tompkins, "Adaptive matched filtering for QRS detection," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Nov. 1998, pp. 147–148.

[7] P. S. Hamilton and W. J. Tompkins, "Quantitative investigation of QRS detection rules using the MIT/BIH arrhythmia database," *IEEE Trans. Biomed. Eng.*, vol. BME-33, no. 12, pp. 1157–1165, Dec. 1986.

[8] J. P. V. Madeiro, P. C. Cortez, J. A. L. Marques, C. R. V. Seisdedos, and C. R. M. R. Sobrinho, "An innovative approach of QRS segmentation based on first-derivative, Hilbert and Wavelet Transforms," *Med. Eng. Phys.*, vol. 34, pp. 1236–1246, Nov. 2012.

[9] P. E. Trahanias, "An approach to QRS complex detection using mathematical morphology," *IEEE Trans. Biomed. Eng.*, vol. 40, no. 2, pp. 201–205, Feb. 1993.

[10] K. Friganovic, A. Jovic, D. Kukolja, M. Cifrek, and G. Krstacic, "Optimizing the detection of characteristic waves in ECG based on exploration of processing steps combinations," in *Proc. IFMBE* 2017, pp. 928–931.

[11] G. Vijaya, V. Kumar, and H. K. Verma, "ANN-based QRS-complex analysis of ECG," *J. Med. Eng. Technol.*, vol. 22, no. 4, pp. 160–167, 1998.

[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[13] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258.

[14] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG heartbeat classification: A deep transferable representation," in *Proc. IEEE Int. Conf. Healthcare Inform. (ICHI)*, Jun. 2018, pp. 443–444.

[15] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.

[16] S. Kiranyaz, T. Ince, R. Hamila, and M. Gabbouj, "Convolutional Neural Networks for patient-specific ECG classification," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Milan, Italy, Aug. 2015, pp. 2608–2611. [Online]. Available: http://ieeexplore.ieee.org/document/7318926/

[17] A. Isin and S. Ozdalili, "Cardiac arrhythmia detection using deep learning," *Procedia Comput. Sci.*, vol. 120, no. 120, pp. 268–275, 2017.

[18] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network," *Inf. Sci.*, vol. 405, pp. 81–90, Sep. 2017.

[19] U. R. Acharya, H. Fujita, S. L. Oh, Y. Hagiwara, J. H. Tan, and M. Adam, "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," *Inf. Sci.*, vols. 415–416, pp. 190–198, Nov. 2017.

[20] U. R. Acharya, H. Fujita, O. S. Lih, M. Adam, J. H. Tan, and C. K. Chua, "Automated detection of coronary artery disease using different durations of ECG segments with convolutional neural network," *Knowl.-Based Syst.*, vol. 132, pp. 62–71, Sep. 2017.

[21] B. S. Chandra, C. S. Sastry, and S. Jana, "Robust heartbeat detection from multimodal data via CNN-based generalizable information fusion," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 3, pp. 710–717, Mar. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8410035/

[22] Y. Xiang, Z. Lin, and J. Meng, "Automatic QRS complex detection using two-level convolutional neural network," *Biomed. Eng. Online*, vol. 17, no. 1, p. 13, 2018.

[23] W. Zhong, L. Liao, X. Guo, and G. Wang, "A deep learning approach for fetal QRS complex detection," *Physiol. Meas.*, vol. 39, no. 4, 2018, Art. no. 045004.

[24] J. S. Lee, M. Seo, S. W. Kim, and M. Choi, "Fetal QRS detection based on convolutional neural networks in noninvasive fetal electrocardiogram," in *Proc. 4th Int. Conf. Frontiers Signal Process. (ICFSP)*, vol. 4, Sep. 2018, pp. 75–78.

[25] R. Yu, Y. Gao, X. Duan, T. Zhu, Z. Wang, and B. Jiao, "QRS detection and measurement method of ECG paper based on convolutional neural networks," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBS)*, Jul. 2018, pp. 4636–4639.

[26] M. Šarlija, F. Jurišić, and S. Popović, "A convolutional neural network based approach to QRS detection," in *Proc. Int. Symp. Image Signal Process. Anal. (ISPA)*, Sep. 2017, pp. 121–125.

[27] J. Malik, Y. L. Lo, and H. T. Wu, "Sleep-wake classification via quantifying heart rate variability by convolutional neural network," *Physiol. Meas.*, vol. 39, no. 8, 2018, Art. no. 085004.

[28] *Recommended Practice for Testing and Reporting Performance Results of Ventricular Arrhythmia Detection Algorithms*, Assoc. Adv. Med. Instrum., Arlington, VA, USA, 1987.

[29] D. M. Pelt and J. A. Sethian, "A mixed-scale dense convolutional neural network for image analysis," *Proc. Nat. Acad. Sci. USA*, vol. 115, no. 2, pp. 254–259, 2018. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1715832114

[30] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale convolutional neural networks for lung nodule classification," in *Information Processing in Medical Imaging* (Lecture Notes in Computer Science), S. Ourselin, D. C. Alexander, C.-F. Westin, and M. J. Cardoso, Eds. Cham, Switzerland: Springer, 2015, pp. 588–599.

[31] J. Hu, Z. Chen, M. Yang, R. Zhang, and Y. Cui, "A multiscale fusion convolutional neural network for plant leaf recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 6, pp. 853–857, Jun. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8302944/

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[33] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.

[34] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, "What is the best multi-stage architecture for object recognition?" in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2146–2153.

[35] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.*, in Lecture Notes in Computer Science, 2014, pp. 834–849.

[36] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 842–850.

[37] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Netw.*, vol. 21, nos. 2–3, pp. 427–436, 2008.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[39] N. Ketkar, "Introduction to PyTorch," in *Deep Learning With Python*. Bangalore, India: Apress, 2017.

[40] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C. K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[41] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, "A database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG," in *Proc. Comput. Cardiol.*, Lund, Sweden, Sep. 1997, pp. 673–676. [Online]. Available: http://ieeexplore.ieee.org/document/648140/

AHSAN HABIB received the B.Sc.Eng. degree in computer science and engineering from the Shahjalal University of Science and Technology, Sylhet, Bangladesh, and the M.Eng. degree in information and communications technologies from the Asian Institute of Technology, Thailand. He is currently pursuing the Ph.D. degree with the School of Information Technology, Deakin University, Australia. His research interests include biomedical signal processing and modeling, time series analysis, machine learning, and deep learning.

CHANDAN KARMAKAR received the B.Sc.Eng. degree in computer science and engineering from the Shahjalal University of Science and Technology, Sylhet, Bangladesh, and the Ph.D. degree from The University of Melbourne, Australia. He joined the School of Information Technology, Deakin University, in 2018, as a Lecturer. He has published one book and more than 130 research articles, including 42 journal articles. His research interests include biomedical devices and signal processing, cardiovascular and neural systems related to sleep-disordered breathing, human gait dysfunctions, cardiovascular diseases, and diabetic autonomic neuropathy.

JOHN YEARWOOD is currently the Head of the School of Information Technology, Deakin University, Australia. His work in data mining and computational intelligence has led to the development of new machine learning and hybrid learning algorithms for artificial neural networks, as well as new data and text mining and pattern recognition approaches. His work in decision science has developed the use of argumentation structures for the modeling of knowledge and collaborative decision making in complex domains. He has published over 200 journal and refereed conference papers, including 2 books. He is the Editor-in-Chief of the *Journal of Research and Practice in Information Technology*.