Deep Neural Network for Respiratory Sound Classification in Wearable Devices Enabled by Patient Specific Model Tuning

Jyotibdha Acharya ⃝, *Student Member, IEEE*, and Arindam Basu ⃝, *Senior Member, IEEE*

Abstract—**The primary objective of this paper is to build classification models and strategies to identify breathing sound anomalies (wheeze, crackle) for automated diagnosis of respiratory and pulmonary diseases. In this work we propose a deep CNN-RNN model that classifies respiratory sounds based on Mel-spectrograms. We also implement a patient specific model tuning strategy that first screens respiratory patients and then builds patient specific classification models using limited patient data for reliable anomaly detection. Moreover, we devise a local log quantization strategy for model weights to reduce the memory footprint for deployment in memory constrained systems such as wearable devices. The proposed hybrid CNN-RNN model achieves a score of 66.31% on four-class classification of breathing cycles for ICBHI'17 scientific challenge respiratory sound database. When the model is re-trained with patient specific data, it produces a score of 71.81% for leave-one-out validation. The proposed weight quantization technique achieves ≈4× reduction in total memory cost without loss of performance. The main contribution of the paper is as follows: Firstly, the proposed model is able to achieve state of the art score on the ICBHI'17 dataset. Secondly, deep learning models are shown to successfully learn domain specific knowledge when pre-trained with breathing data and produce significantly superior performance compared to generalized models. Finally, local log quantization of trained weights is shown to be able to reduce the memory requirement significantly. This type of patient-specific re-training strategy can be very useful in developing reliable long-term automated patient monitoring systems particularly in wearable healthcare solutions.**

Index Terms—**CNN, LSTM, patient specific model, respiratory audio analysis, weight quantization.**

I. INTRODUCTION

TWO most clinically significant lung sound anomalies are wheeze and crackle. Wheeze is a continuous high pitched adventitious sound that results from obstruction of breathing airway. While normal breathing sounds have majority of their energy concentrated in 80–1600 Hz [1], wheeze sounds have been shown to be present in the frequency range 100 Hz-2 KHz. Wheeze is normally associated with patients suffering from asthma, chronic obstructive pulmonary disease (COPD) etc. Crackles are explosive and discontinuous sounds present during inspiratory and expiratory parts of breathing cycle with a significantly smaller duration compared to the total breathing cycle. Crackles have been associated with obstructive airway diseases and interstitial lung diseases [2].

Auscultation has been used historically for screening and monitoring respiratory diseases. It provides a simple and non-invasive approach to detect respiratory and cardiovascular diseases based on lung sound abnormalities. But these methods suffer from two disadvantages. Firstly, a trained medical professional is required to diagnose a patient based on adventitious lung sounds and therefore, disproportionate number of medical practitioners compared to overall population hinders the speed at which patients are tested. Secondly, even if the patients are diagnosed by experienced professionals, there might be subjectivity in the diagnosis due to dissimilar interpretation of the respiratory sounds by different medical professionals [3].

So, in the past decade several attempts were made to design algorithms and feature extraction techniques for automated detection of breathing anomalies. Some popular feature extraction techniques used include spectrogram [4], Mel-Frequency Cepstral Coefficients (MFCC) [5], wavelet coefficients [6], entropy based features [7] etc. Several machine learning (ML) algorithms have been developed in past few years to detect breathing sound anomalies such as logistic regression [8], Dynamic Time Wrap (DTW), Gaussian mixture model (GMM) [9], random forest [4], Hidden Markov Model (HMM) [10] etc. An exploration of existing literature reveals some conspicuous issues with these approaches. Firstly, most of the ML algorithms use manually crafted highly complex features suitable for their algorithms and due to absence of publicly available datasets, it was hard to compare the efficacy of the feature extraction methods and algorithms proposed [11]. Secondly, most of the strategies were developed for a binary classification problem to identify either wheeze or crackle and therefore, not suitable for multi-class classification to detect wheeze and crackle simultaneously [12]. These drawbacks make these approaches difficult to apply in real world scenarios.

Deep learning has gained a lot of attention in recent years due to its unparalleled success in a variety of applications including clinical diagnostics and biomedical engineering [13].

A significant advantage of these deep learning paradigms is that there is no need to manually craft features from the data since the network learns useful features and abstract representations from the data through training. Due to wide success of convolutional neural networks (CNN) in image related tasks, they have been extensively used in biomedical research for image classification [14], anomaly detection [15], image segmentation [16], image enhancement [17], automated report generation [18] etc. There have been multiple successful applications of deep CNNs in diagnosis of cardiac diseases [19], neurological diseases [20], cancer [21] and ophthalmic diseases [22]. While CNNs have shown significant promise for analyzing image data, recurrent neural networks (RNN) are better suited for learning long term dependencies in sequential and time-series data [23]. The state of the art systems in natural language processing (NLP), audio and speech processing use deep RNNs to learn sequential and temporal features [24]. Finally, hybrid CNN-RNN models have shown significant success in video analytics [25] and speech recognition [26]. These hybrid models show particular promise in cases where both spatial and temporal/sequential features need to be learned from the data.

Since deep learning came into prominence, it is also being used by researchers for audio based biomedical diagnosis and anomaly detection. Some significant areas of audio based diagnosis using deep learning include sleep apnea detection, cough sound identification, heart sound classification etc. Amoh *et al.* [27] used a chest mounted sensor to collect audio containing both speech and cough sounds and then used both CNN and RNN to identify cough sounds. Similarly, Nakano *et al.* [28] trained a deep neural networks on tracheal sound spectrograms to detect sleep apnea. In [29], authors train a CNN architecture to classify heart sounds into normal and abnormal classes. The non-invasive nature of audio based diagnosis make them an attractive choice for biomedical applications.

A major handicap in training a deep network is that a significantly large dataset and considerable time and resources need to be allocated for the training. While the second issue can be solved by using dedicated deep learning accelerators (GPU,TPU etc), the first issue is even more exacerbated for medical research since medical datasets are very sparse and difficult to obtain. One way to circumvent this issue is to use transfer learning. The central idea behind transfer learning is following: a deep network trained in a domain $D1$ to perform task $T1$ can successfully use the learned data representations to perform task $T2$ in domain $D2$. Most commonly used method for transfer learning is using a large dataset to train a deep network and then re-training a small section of the network on the available data (often significantly small) for the specific task and specific domain. Transfer learning has been used in medical research for cancer diagnosis [30], prediction of neurological diseases [31] etc. While traditionally transfer learning refers to transfer of knowledge between two disparate domains, for biomedical research, it is also used for knowledge transfer in the same domain where a model is trained on a larger population dataset and the knowledge is then transferred for context specific learning on a smaller dataset [32], [33]. This strategy is specially useful for biomedical applications due to scarcity of domain specific patient data.

Finally, for employing machine learning methods for medical diagnosis, two primary approaches are used. The first one is generalized models where the models are trained on a database of multiple patient data and it is tested on new patient data. This type of models learns generalized features present across all the patients. While this kind of models are often easier to deploy, they often suffer from inter-patient variability of features and may not produce reliable results for unseen patient data. The second approach is patient specific models, where the models are trained on patient specific data to produce more precise results for the patient specific diagnosis. While these models are harder to train due to difficulty in collecting large amount of patient specific data, they often produce very reliable and consistent results [34]. While a patient specific model requires additional time and effort from healthcare professionals for collecting and labeling the data, specially for chronic diseases where long term monitoring is of the essence, this additional effort is well compensated by reduced hospitalization and reduced loss of time from work for the patient resulting from better continuous monitoring [35].

Since a large fraction of medical diagnosis algorithms are geared toward wearable devices and mobile platforms, large memory and computational power requirement of deep learning methods present a considerable challenge for commercial deployment. Weight quantization [7], low precision computation [36] and lightweight networks [37] are some of the approaches used to address this challenge. Quantizing the weights of the trained network is the most straight-forward way to reduce the memory requirement for deployment. DNNs with 8 or 16 bit weights have been shown to achieve comparable accuracy compared to their full precision counterpart [7]. Though linear quantization is most commonly used, log quantization has been shown to achieve similar accuracy at lower bit precision [38]. Finally, lightweight networks like MobileNet [37] reduces computational complexity and memory requirement without significant loss of accuracy by replacing traditional convolution layers by depthwise separable convolution.

In this paper we propose a hybrid CNN-RNN model to perform four class classification of breathing sounds on International Conference on Biomedical and Health Informatics (ICBHI'17) scientific challenge respiratory sound database [39] and then devise a screen and model tuning strategy to build patient specific diagnosis models from limited patient data. For comparison of our model with more commonly used CNN architectures, we applied the same methodology on VGGnet [40] and Mobilenet [37] architecture. Finally, we propose a layerwise logarithmic quantization scheme that can reduce the memory footprint of the networks without significant loss of performance. The sections are organized as follows: Section II describes the dataset, feature extraction method, proposed deep learning model and weight quantization. Section III tabulates the results for generalized and patient specific model and quantization performance. Finally, Section IV discusses the conclusions and main contributions of the paper.

II. MATERIALS AND METHODS

A. Dataset

For this work we have used the International Conference on Biomedical and Health Informatics (ICBHI'17) scientific challenge respiratory sound database [39]. This is the largest publicly available respiratory sound database. The database contains 920 recordings from 126 patients. Each breathing cycle in a recording is annotated by respiratory experts as one of the four classes: normal, wheeze, crackle and both (wheeze and crackle). The database contains a total of 6898 respiratory cycles out of which 1864 cycles contain crackles, 886 contain wheeze, 506 contain both and rest are normal. The dataset contains samples recorded with different equipment (AKG C417 L Microphone, 3 M Littmann Classic II SE Stethoscope, 3 M Litmmann 3200 Electronic Stethoscope and WelchAllyn Meditron Master Elite Electronic Stethoscope) from hospitals in Portugal and Greece. The data is recorded from different locations of chest: 1) Trachea 2) Anterior left 3) Anterior right 4) Posterior left 5) Posterior right 6) Lateral left and 7) Lateral right. Furthermore, a significant number of samples are noisy. These characteristics make the classification problem more challenging and much closer to real world scenarios compared to manually curated datasets recorded under ideal conditions. Further details about the database and data collection methods can be found in [39].

B. Evaluation Metrics

In the original challenge, out of 920 recordings, 539 recordings were marked as training samples and 381 recordings were marked as testing samples. There are no common patients between training and testing set. For this work we used the officially described evaluation metrics for the four-class (normal(N), crackle(C), wheeze(W) and both(B)) classification problem defined as follows:

$$Sensitivity(Se) = \frac{C_{correct} + W_{correct} + B_{correct}}{C_{total} + W_{total} + B_{total}} \quad (1)$$

$$Specificity(Sp) = \frac{N_{correct}}{N_{total}} \quad (2)$$

$$Score(Sc) = \frac{Se + Sp}{2} \quad (3)$$

where $i_{correct}$ and i_{total} represent correctly classified and total breathing cycles of $class_i$ respectively. Since deep learning models require a large amount of data for training, we use a 80–20 split of patients for training and testing in all our experiments.The training set contains recordings from 101 patients while testing set contains recordings from 25 patients.

For a more complete evaluation of the proposed model, we also evaluate it using other commonly used metrics such as precision, recall and f1-score. Moreover, the dataset has disproportionate number of normal vs anomalous samples and the official metrics are micro-averaged (calculated over all the classes).Therefore, there is a chance that performance of the models on one class overshadows the other classes in overall

results. Therefore, we calculated the precision, recall and f1-score using macro-averaging (metrics are computed for each class individually and then averaged).

C. Related Work

A number of papers have been published so far analyzing this dataset. Jakovljevic et al. [10] used hidden markov model with Gaussian mixture model to classify the breathing cycles. They have used spectral subtraction based noise suppression to preprocess the data and MFCC features are used for classification. Their models obtained a score of 39.56% on the original train-test split and 49.5% on 10-fold cross-validation of the training set.

Kochetov et al. [41] proposed a noise marking RNN for the four-class classification. Their proposed model contains two sections: an attention network for binary classification of respiratory cycles into noisy and non-noisy classes and an RNN for four class classification. The attention network learns to identify noisy parts of the audio and suppress those sections and passes the filtered audio to the RNN for classifications. With a 80–20 split, they obtained a score of 65.7%. They didn't report the score for the original train-test split. Though this method reports relatively higher scores, one primary issue with this method is that there are no noise labels in the metadata of the ICBHI dataset and the paper doesn't mention any method for obtaining these labels. Since there are no known objective methods to measure the noise labels in these type of audio signals, this kind of manual labeling of the respiratory cycles makes their results unreliable and irreproducible.

Perna et al. [42] used a deep CNN architecture to classify the breathing cycles into healthy and unhealthy and obtained an accuracy of 83% using a 80–20 train-test split and MFCC features. They also did a ternary classification of the recordings into healthy,chronic and non-chronic diseases and obtained an accuracy of 82%.

Chen et al. [12] used optimized S-transform based feature maps along with deep residual nets (ResNets) on a smaller subset of the dataset (489 recordings) to classify the samples (not individual breathing cycles) into three classes (N, C and W) and obtained an accuracy of 98.79% on a 70–30 train-test split.

Finally, Chambres et al. [43] have proposed a patient level model where they classify the individual breathing cycles into one of the four classes using lowlevel features (melbands, mfcc, etc), rythm features (loudness, bpm etc), the SFX features (harmonicity and inharmonicity information) and the tonal features (chords strength, tuning frequency etc). They used boosted tree method for the classification. Next, they classified the patients as healthy or unhealthy based on the percentage of breathing cycles of the patient classified as abnormal. They have obtained an accuracy of 49.63% on the breathing cycle level classification and an accuracy of 85% on patient level classification. The justification for this patient level model is that medical professionals do not take decisions about patients based on individual breathing cycles but rather based on longer breathing sound segments and the trends represented by several breathing cycles

TABLE I
SUMMARY OF EXISTING LITERATURE ON ICBHI DATASET

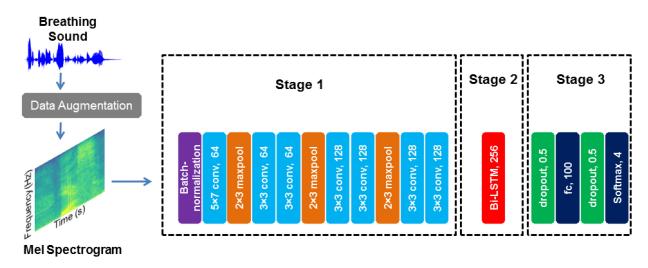| Paper	Features	Classification Method	Results
Jakovljevic *et al.* [10]	MFCC	GMM+ HMM	sc: 39.56% (original train test split), 49.5% (training data, 10-fold cross-validation)
Kochetov *et al.* [41]	MFCC	Noise masking RNN	sc: 65.7% (80 − 20 split, four class classification)
Perna *et al.* [42]	MFCC	CNN	Acc: 83% (80 − 20 split, healthy-unhealthy classification), Acc: 82% (healthy, chronic and non-chronic classification)
Chen *et al.* [12]	optimized S-transform	ResNets	sc: 98.79% (smaller subset of original data, 70 − 30 split, sample level classification)
Chambres *et al.* [43]	Multiple features	boosted tree	sc: 49.63% (original train test split), Acc: 85% (original train test split, patient level healthy-unhealthy classification)

Fig. 1. Hybrid CNN-RNN: a three stage deep learning model. Stage 1 is a CNN that extracts abstract feature maps from input Mel-spectrograms, stage 2 consists of a Bi-LSTM layer that learns temporal features and stage 3 consists of fully connected (FC) and softmax layers that convert outputs to class predictions.

of a patient can provide a more consistent diagnosis. A summary of the literature is presented in Table I.

D. Proposed Method

1) Feature Extraction and Data Augmentation: Since the audio samples in the dataset had different sampling frequencies, first all of the signals were downsampled to 4 kHz. Since both wheeze and crackle signals are typically present within frequency range 0–2 kHz, downsampling the audio samples to 4 kHz should not cause any loss of relevant information.

As the dataset is relatively small for training a deep learning model, we used several data augmentation techniques to increase the size of the dataset. We used noise addition, speed variation, random shifting, pitch shift etc to create augmented samples. Aside from increasing the dataset size, these data augmentation methods also help the network learn useful data representations in-spite of different recording conditions, different equipments, patient age and gender, inter-patient variability of breathing rate etc.

For feature extraction we have used Mel-frequency spectrogram with a window size of 60 ms with 50% overlap. Each breathing cycle is converted to a 2D image where rows correspond to frequencies in Mel scale and columns correspond to time (window) and each value represent log amplitude value of the signal corresponding to that frequency and time window.

2) Hybrid CNN-RNN: We propose a hybrid CNN-RNN model (Fig. 1) that consists of three stages: the first stage is a deep CNN model that extracts abstract feature representations from the input data, the second stage consists of a bidirectional long short term memory layer (Bi-LSTM) that learns temporal relations and finally in the third stage we have fully connected and softmax layers that convert the output of previous layers to class prediction. While these type of hybrid CNN-RNN architectures have been more commonly used in sound event detection ([44], [45]), due to sporadic nature of wheeze and crackle as well as their temporal and frequency variance, similar hybrid architectures may prove useful for lung sound classification.

The first stage consists of batch-normalization, convolution and max-pool layers. The batch normalization layer scales the input images over each batch to stabilize the training. In the 2D convolution layer the input is convolved with 2D kernels to produce abstract feature maps. Each convolution layer is followed by Rectified Linear activation functions (ReLU). The max-pool layer selects the maximum values from a pixel neighborhood which reduces the overall network parameters and results in shift-invariance [13].

LSTM have been proposed by Hochreiter and Schmidhuber [46] consisting of gated recurrent cells that block or pass the data in a sequence or time series by learning the perceived importance of data points. Each current output and the hidden state of a cell is a function of current as well as past values of the

data. Bidirectional LSTM consists of two interconnected LSTM layers, one of which operates on the same direction as data sequence while the other operates on the reverse direction. So, the current output of the Bi-LSTM layer is function of current, past and future values of the data. We used tanh as non-linear activation function for this layer.

The final fully connected and softmax layers take the output of the Bi-LSTM layer and convert it to class probabilities $p_{class} \in [0, 1]$. Finally, the model is trained with categorical crossentropy loss and Adam optimizer for the four class classification problem. We also used dropout regularization in the fully connected layer to reduce overfitting. The parameters for model architecture (such as number of layers, number of filters etc.) were optimized based on minimizing training loss. Finally, to handle the class imbalance pproblem in the dataset, higher costs were assigned to classes with smaller number of samples during model training.

To benchmark the performance of our proposed model, we compare it to two standard CNN models, VGG-16 [40] and Mobilenet [37]. Since our dataset size is limited even after data augmentation, it can cause overfitting if we train these models from scratch on our dataset. Hence, we used Imagenet trained weights instead and replaced the dense layers of these models with an architecture similar to the fully connected and softmax layers of our proposed CNN-RNN architecture. Then the models are trained with a small learning rate.

3) Patient Specific Model Tuning: Though most of the existing research concentrate on developing generalized models for classifying respiratory anomalies, the problem with this kind of models is that their performance can often deteriorate for a completely new patient due to inter-patient variability. This kind of inconsistent performance of classification models make them unreliable and thus hinders their wide scale adoption. To qualitatively evaluate the inter-patient variability, we show the boxplot of inter-patient and intra-patient variability of a diverse set of audio features (duration, RMSE, bandwidth, Roll-off, ZCR) in Fig. 2. For the intra-patient variability, we normalized each audio feature of a sample by average of that feature for the samples from that specific patient while for inter-patient variability, we normalized the audio features by the average of that feature over the entire dataset. As evident from the figure, inter-patient variability is significantly larger when compared to intra-patient variability.

Also, from a medical professional's perspective, for most of the chronic respiratory patients, some patient data is already available or can be collected and automated long-term monitoring of patient condition after initial treatment is often very important. Though training a model based on existing patient specific data to extract patient specific features result in a more consistent and reliable patient specific model, it is often very difficult to collect enough data from a patient to sufficiently train a machine learning model. Since deep learning models require much larger amount of data for training, the issue is further exacerbated.

To address these shortcomings of existing methods, we propose a patient specific model tuning strategy that can take advantage of deep learning techniques even with small amount of
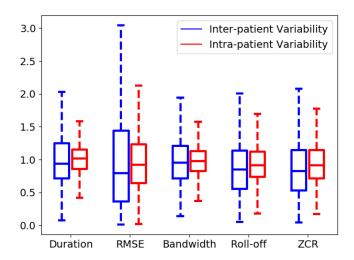

Fig. 2. Boxplot of intra-patient and inter-patient variability of audio features: Intra-patient variability is computed by normalizing each audio feature by average of that feature for the corresponding patient while for Inter-patient variability, the normalization is done by average of the audio feature over the entire dataset. Diverse set of features are used for comparison including breathig cycle duration, energy related feature (RMS ennergy), noise related feature (ZCR) and spectral features (bandwidth, roll-off). Inter-patient variability is significantly larger than intra-patient variability for all the cases.

patient data available. In this proposed model, the deep network is first trained on a large database to learn domain specific feature representations. Then a smaller part of the network is re-trained on the small amount of patient specific data available. This enables us to transfer the learned domain specific knowledge of the deep network to patient specific models and thus produce consistent patient specific class predictions with high accuracy. In our proposed model we train the 3 stage network on the training samples. Then, for a new patient, only the last stage is re-trained with patient specific breathing cycles while the learned CNN-RNN stage weights are frozen in their pre-trained values. For our proposed strategy only ∼1.4% of the network parameters are retrained for patient specific models. For VGG-16 and MobileNet, the same strategy is applied.

4) Weight Quantization: In this proposed weight quantization scheme, the magnitude of weights of each layer are quantized in log domain. The quantized weight ($\tilde{w}$) can be represented as:

$$\tilde{w} = \left\lfloor 2^N \times \frac{w_{\log} - w_{\log}^{\min}}{w_{\log}^{\max} - w_{\log}^{\min}} \right\rfloor \qquad (4)$$

where $w_{\log}$ represents the weights (w) mapped to log domain ($\log_{10}(w)$) and N is the bit precision. The total number of bits required to store each weight in this scheme is (N + 1) since one bit is required to store the sign of the bit. Now, the minimum and maximum weights ($w_{\log}^{\min}$ and $w_{\log}^{\max}$) used for normalization can be calculated globally (over the entire network) or locally (for each layer). Since the architectures used here have different types of layers (convolution, batchnormalization, LSTM etc.) which often show different ranges of weights [38], local weight normaliztion seems to be more logical choice. While local normalization requires minimum and maximum weights of each

TABLE II
COMPARISON OF RESULTS

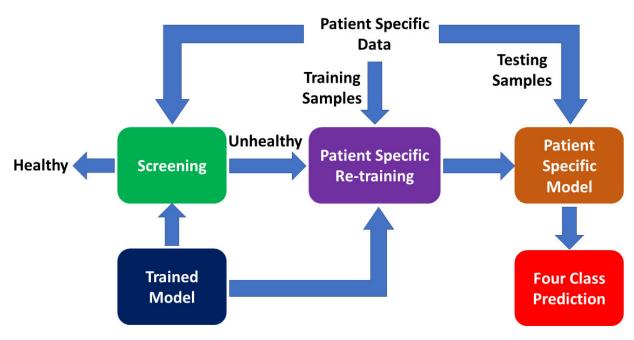| Model	Micro Metrics			Macro Metrics		
	Sensitivity	Specificity	Score	Precision	Recall	f1-score
VGG-16	40.63%	86.03%	63.33%	49.09%	49.24%	48.99%
Mobilenet	46.31%	78.20%	62.26%	52.02%	55.21%	53.13%
Our work	48.63%	84.14%	66.38%	58.47%	58.01%	57.91%

Fig. 3. Screen and model tuning strategy: First the patients are screened into healthy and unhealthy based on % of breathing cycles predicted as unhealthy. For patients predicted to be unhealthy, trained model is re-trained on patient specific data to produce patient specific model which then performs the four class prediction on breathing cycles.

layer to be saved in memory for retrieving the actual weights, this is very insignificant compared to total memory required to save the quantized weights. Finally, we rounded very small weights to zero before applying log quantization to limit the quantization range in log domain.

III. RESULTS AND DISCUSSIONS

A. Generalized Model

Firstly, we evaluated our model on four class breathing cycle using micro and macro metrics described in Section II-B. To compare our results with traditionally used CNN architectures, we used VGGnet and Mobilenet. All models were trained and tested on a workstation with Intel Xeon E5-2630 CPU with 128 GB RAM and NVIDIA TITAN Xp GPU. The results are tabulated in Table II. The results are averaged over five randomized train-test sets. As it can be seen from the Table II, the proposed hybrid CNN-RNN model trained with data augmentation produces state of the art results. Both VGG-16 and Mobilenet produces slightly lower scores both in terms of macro and micro metrics. The score obtained by the proposed model also outperform results reported by Kochetov *et al.* using noise labels on a similar 80–20 split (Table I). We have also performed a 10-fold cross-validation on the dataset for our proposed model

and the average score obtained is 66.43%. Due to unavailability of similar audio datasets in biomedical field, we have also tested the proposed hybrid model on Tensorflow speech recognition challenge [47] to benchmark its performance. For an eleven-class classification with $90\% - 10\%$ train-test split, it produced a respectable accuracy of 96%. For the sake of completeness, we also tested the dataset using same train-test split strategy with a variety of commonly used temporal and spectral features (RMSE, ZCR, spectral centroid, roll-off frequency, entropy, spectral contrast etc. [48]) with non-DL methods such as SVM, shallow neural network, random forest and gradient boosting. The resulting scores were significantly lower (44.5–51.2%).

B. Patient Specific Model Tuning Strategy

It has been shown by Chambres *et al.* [43] that though it is difficult to achieve high scores for breathing cycle level classification, it is much easier to achieve high accuracy in patient level binary classification (healthy/sick). Hence, we propose a screen and model tuning strategy. First, the patients are screened using the pre-trained model and if a patient is found to be unhealthy, the pre-trained model is retrained on the patient data to build the patient specific model that can monitor the patient condition in future with higher reliability. The proposed model is shown in Fig. 3. To evaluate the performance of the proposed

TABLE III
COMPARISON OF PATIENT SPECIFIC MODELS

Pre-trained Model	Patient Specific Training	Se	Sp	Sc
NA.	Majority Class	46.25%	80.45%	63.35%
VGG-16 (ImageNet)	SVM	50.35%	82.38%	66.36%
VGG-16 (ICBHI)	Dense + Softmax	54.41%	82.66%	68.54%
MobileNet (ICBHI)	Dense + Softmax	51.28%	83.92%	67.60%
Hybrid CNN-RNN (TF speech dataset)	Dense + Softmax	43.01%	74.28%	58.4%
Hybrid CNN-RNN (UrbanSound8K dataset)	Dense + Softmax	34.73%	88.45%	61.59%
Hybrid CNN-RNN (ICBHI dataset)	Dense + Softmax	56.91%	86.70%	71.81%

methodology, we used leave one out validation. Since there are variable number of recordings from each patient in the dataset, of the n samples from a patient, $n-1$ samples are used to retrain the model and it is tested on the other sample. This method is repeated so that all the samples are in the test set once. We trained the proposed model on the patients in the train set and evaluated it on the patients in test set. Since leave one out validation is not possible for patients with only one sample, we only considered patients with more than one sample. The dataset contains different number of recordings from each patient and the length of the recordings and number of breathing cycles in each recording vary widely. But on average, ≈ 47 patient breathing cycles are used for the fine-tuning of patient specific models.

Now, for an objective evaluation of the proposed method, we need to consider a few things. Firstly, if for a patient, most of the breathing cycles belong to one of the classes, the simplest strategy is to assign that class to any of the breathing cycles in the test set. We define this strategy as majority class and use it as a baseline to evaluate the performance of our model. The predicted class probability of kth class is defined as:

$$p_{class^k} = 1 \quad \text{if} \quad N_B^k > N_B^i \quad \forall \quad i \neq k$$
$$= 0 \quad \text{otherwise} \tag{5}$$

where N_B^i is the number of breathing cycles belonging to class i to the specific patient.

Secondly, since we are using patient specific data to train the models, we have to verify if our proposed model tuning strategy provides any advantage over a simple classifier trained on only patient specific data. To verify this, we used an ImageNet [49] trained VGG-16 [40] as a feature extractor along with an SVM classifier to build patient specific models. Variants of VGG trained on ImageNet dataset have been shown to be very efficient feature extractors not only for image classification, but also for audio classification [50]. Here we use the pre-trained CNN to extract features from patient recordings and train an SVM based on those features only on the patient specific data.

Thirdly, we are proposing that by pre-training the hybrid CNN-RNN model on the respiratory data, the model learns domain specific feature representations that are transferred to the patient specific model. To justify this claim, we trained the same model on tensorflow speech recognition challenge dataset [47] as well as urban sounds 8 K dataset [51]. Then we used the same model tuning strategy to re-train the model on patient specific data. If the proposed model learns only the audio feature specific abstract representations from the data, then a model trained on any sufficiently large audio database should perform
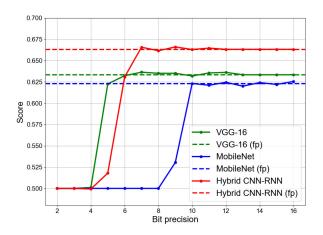

Fig. 4. Local log quantization: Score achieved by VGG-16, MobileNet and hybrid CNN-RNN with varying bit precision under local log quantization. VGG-16 requires minimum bit precision to achieve full precision (fp) accuracy while MobileNet requires maximum bit precision.

well. But, if the model learns respiratory sound domain specific features from the data, the model pre-trained on respiratory sounds should outperform the model pre-trained on any other type of audio database. Finally, we comapare the results of our model with pure CNN models VGG-16 and MobileNet using the same experimental methodology.

The results are tabulated in Table III. Firstly, Our proposed strategy outperforms all other models and strategies and obtains a score of 71.81%. Secorndly, VGG-16 and MobileNet achieves scores 68.54% and 67.60% which signifies pure CNNs can be employed for respiratory audio classification, albeit not as effective as a CNN-RNN hybrid model. Thirdly, results corresponding to both audio trained networks shows that audio domain pre-training is not very effective for respiratory domain feature extraction. We explain this observation in further details in Section IV. Finally, Imagenet trained VGG-16 shows promise as a feature extractor for respiratory data, although it does not reach the same level of performance as ICBHI trained models.

C. Memory and Computational Complexity

Even though the proposed models show excellent performance in the classification task, the memory requirement for storing huge number of weights for these models make them unsustainable for application in mobile and wearable platforms. Hence, we apply the local log quantization scheme proposed in Section II-D4. Fig. 4 shows the score achieved by the models as a function of bit precision of weights. As expected, VGG-16
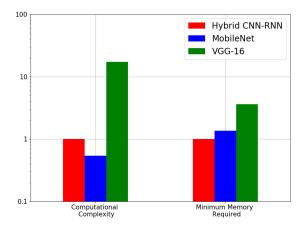
Fig. 5. Resource comparison: Comparison of normalized computational complexity (GFLOPS/sample) and minimum memory required (Mbits) by VGG-16, MobileNet and hybrid CNN-RNN. MobileNet and hybrid CNN-RNN present a trade-off between computational complexity and memory required for optimum performance.

outperforms the other to models due to its over-parameterized design [38]. MobileNet shows particularly poor performance in weight quantization and is only able to achieve optimum accuracy at 10 bit precision. This poor quantization performance can be attributed to large number of batch-normalization layers and RELU6 activation of MobileNet architecture [38]. While several approaches have been proposed to circumvent these issues [52], these methods are not compatible with Imagenet pre-trained MobileNet model since they focus on modifications in the architecture rather than quantization of pre-trained weights. The hybrid CNN-RNN model performs slightly worse than VGG-16 since it has LSTM layer which requires higher bit precision compared to the CNN counterpart [53].

Finally, we compare the computational complexity and minimum memory requirement of the models. The computational complexity is calculated in GFLOPS per sample during inference. The minimum memory required for a model is calculated as (total number of parameters $\times N^{opt}$) where N^{opt} is the minimum bit precision required for the model to achieve optimum accuracy. The only other major source of computational complexity for end-to-end classification of respiratory audio is computations required for converting audio samples to their respective mel-spectrograms. Since this compuatational overhead for feature extraction is significantly less (<1 MFLOPS) [4] compared to the DL architecture, we can ignore it in computational complexity calculations. Both computational complexity and memory are normalized with respect to hybrid CNN-RNN model and shown in Fig. 5. Due to large number of parameters, even after significant memory compression through weight quantization, VGG-16 requires significantly higher memory. Hybrid CNN-RNN model requires least amount of memory followed by MobileNet. In terms of computational complexity, while VGG-16 requires very high computational overhead, MobileNet is computationally most efficient followed by hybrid CNN-RNN. Hence, MobileNet is more suitable for a hardware system where power is the primary constraint while hybrid

CNN-RNN, while less power efficient than MobileNet, achieves better performance at minimal memory footprint.

Finally, Our proposed system requires data pre-processing, feature extraction and classification only once in each breathing cycle. Therefore, if we consider a ping-pong buffer architecture [54] for audio acquisition and processing, our system needs to perform end to end classification of breathing cycles at a latency smaller than minimum breathing cycle duration for real time operation. The primary computational bottleneck of the proposed system is the DL architecture as mentioned earlier. The number of computations of the proposed architecture is of the same order as Mobilenet as shown in Fig. 5. Since the minimum breathing cycle duration is >1 second [55] and the per sample latency of Mobilenet on modern mobile SoCs is only ~100 ms [56], the proposed system should easily be able to perform real time classification of respiratory anomalies.

IV. CONCLUSION

In this paper, we have developed a hybrid CNN-RNN model that produces state of the art results for ICBHI'17 respiratory audio dataset. It produces a score of 66.31% score on 80–20 split for four-class respiratory cycle classification. We also propose a patient screening and model tuning strategy to identify unhealthy patients and then build patient specific models through patient speecific re-training. This proposed model provides significantly more reliable results for the original train-test split achieving a score of 71.81% for leave-one-out cross-validation. It is observed that trained models from image recognition field, surprisingly, perform better in transferring knowledge than those pre-trained on speech. A possible explanation for this could be that while image-trained models are trained on a much larger Imagenet dataset and therefore, has better generalization performance compared to models trained on relatively smaller audio datasets. While lack of availability of pre-trained models in audio domain and prohibitively long training time required for training a model with audio datasets of sizes comparable to Imagenet prevent us from verifying this hypothesis in this work, in future we plan to explore transfer learning performance of audio and image datasets in further detail. We also develop a local log quantization strategy for reducing the memory cost of the models that achieves $\approx4\times$ reduction in minimum memory required without loss of performance. The primary significance of this result is that this weight quantization strategy is able to achieve considerable weight compression without any architectural modification to the model or quantization aware training. Finally, while the proposed model has higher computational complexity than MobileNet, it has minimal memory footprint among the models under consideration. Since the amount of data from a single patient is still very small for this dataset, in future, we plan to employ this strategy with larger amount of patient specific data. We also plan to create an embedded implementation of this algorithm for a wearable device to be used in patient monitoring at home. Further reductions in computational complexity will be explored using a neuromorphic spike based approach [57], [58].

References

[1] N. Gavriely, Y. Palti, G. Alroy, and J. B. Grotberg, "Measurement and theory of wheezing brea th sounds," *J. Appl. Physiol.*, vol. 57, no. 2, pp. 481–492, 1984.

[2] P. Piirila and A. Sovijarvi, "Crackles: recording, analysis and clinical significance," *Eur. Respir. J.*, vol. 8, no. 12, pp. 2139–2148, 1995.

[3] M. Bahoura and C. Pelletier, "Respiratory sounds classification using gaussian mixture models," in *Proc. IEEE Can. Conf. Elect. Comput. Eng. (IEEE Cat. No. 04CH37513)*, 2004, vol. 3, pp. 1309–1312.

[4] J. Acharya, A. Basu, and W. Ser, "Feature extraction techniques for low-power ambulatory wheeze detection wearables," in *Proc. IEEE 39th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2017, pp. 4574–4577.

[5] B.-S. Lin and B.-S. Lin, "Automatic wheezing detection using speech recognition technique," *J. Med. Biol. Eng.*, vol. 36, no. 4, pp. 545–554, 2016.

[6] M. Bahoura, "Pattern recognition methods applied to respiratory sounds classification into normal and wheeze classes," *Comput. Biol. Med.*, vol. 39, no. 9, pp. 824–843, 2009.

[7] J. Zhang, W. Ser, J. Yu, and T. Zhang, "A novel wheeze detection method for wearable monitoring systems," in *Proc. IEEE Int. Symp. Intell. Ubiquitous Comput. Edu.*, 2009, pp. 331–334.

[8] P. Bokov, B. Mahut, P. Flaud, and C. Delclaux, "Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population," *Comput. Biol. Med.*, vol. 70, pp. 40–50, 2016.

[9] I. Sen, M. Saraclar, and Y. P. Kahya, "A comparison of svm and gmm-based classifier configurations for diagnostic classification of pulmonary sounds," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 7, pp. 1768–1776, Jul. 2015.

[10] N. Jakovljević and T. Lončar-Turukalo, "Hidden markov model based respiratory sound classification," in *Precision Medicine Powered by pHealth and Connected Health*. Berlin, Germany: Springer, 2018, pp. 39–43.

[11] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PloS One*, vol. 12, no. 5, 2017, Art. no. e0177926.

[12] H. Chen, X. Yuan, Z. Pei, M. Li, and J. Li, "Triple-classification of respiratory sounds using optimized s-transform and deep residual networks," *IEEE Access*, vol. 7, pp. 32 845–32 852, 2019.

[13] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[14] E. Hosseini-Asl *et al.*, "Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network," *Frontiers Biosci.* (Landmark edition), vol. 23, pp. 584–596, 2018.

[15] M. J. van Grinsven, B. van Ginneken, C. B. Hoyng, T. Theelen, and C. I. Sánchez, "Fast convolutional neural network training using selective data sampling: Application to hemorrhage detection in color fundus images," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1273–1284, May 2016.

[16] Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 10, pp. 2421–2433, Oct. 2015.

[17] O. Oktay *et al.*, "Multi-input cardiac image super-resolution using convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2016, pp. 246–254.

[18] P. Kisilev, E. Sason, E. Barkan, and S. Hashoul, "Medical image description using multi-task-loss cnn," in *Deep Learning and Data Labeling for Medical Applications*. Berlin, Germany: Springer, 2016, pp. 121–129.

[19] P. V. Tran, "A fully convolutional neural network for cardiac segmentation in short-axis MRI," *CoRR*, 2016. [Online]. Available: http://arxiv.org/abs/1604.00494

[20] H. K. van der Burgh, R. Schmidt, H.-J. Westeneng, M. A. de Reus, L. H. van den Berg, and M. P. van den Heuvel, "Deep learning predictions of survival based on mri in amyotrophic lateral sclerosis," *NeuroImage: Clin.*, vol. 13, pp. 361–369, 2017.

[21] T. Kooi, B. van Ginneken, N. Karssemeijer, and A. den Heeten, "Discriminating solitary cysts from soft tissue lesions in mammography using a pretrained deep convolutional neural network," *Med. Phys.*, vol. 44, no. 3, pp. 1017–1027, 2017.

[22] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2015, pp. 715–718.

[23] Y. Bengio, P. Simard, P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[24] H. Salehinejad, S. Sankar, J. Barfett, E. Colak, and S. Valaee, "Recent advances in recurrent neural networks," *CoRR*, 2018. [Online]. Available: http://arxiv.org/abs/1801.01078

[25] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional lstm with cnn features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.

[26] Y. Zhao, X. Jin, and X. Hu, "Recurrent convolutional neural network for speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 5300–5304.

[27] J. Amoh and K. Odame, "Deep neural networks for identifying cough sounds," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 5, pp. 1003–1011, Oct. 2016.

[28] H. Nakano, T. Furukawa, and T. Tanigawa, "Tracheal sound analysis using a deep neural network to detect sleep apnea," *J. Clin. Sleep Med.*, vol. 15, no. 08, pp. 1125–1133, 2019.

[29] H. Ryu, J. Park, and H. Shin, "Classification of heart sound recordings using convolution neural network," in *Proc. IEEE Comput. Cardiol. Conf.*, 2016, pp. 1153–1156.

[30] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J.-H. Mao, "Unsupervised transfer learning via multi-scale convolutional sparse coding for biomedical applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1182–1194, May 2018.

[31] A. Payan and G. Montana, "Predicting alzheimer's disease: A neuroimaging study with 3D convolutional neural networks," in *Proc. Int. Conf. Pattern Recognit. App. Methods*, M. D. Marsico, M. A. T. Figueiredo, and A. L. N. Fred, eds., vol. 2, Jan. 2015, pp. 355–362.

[32] L. S. Hu *et al.*, "Accurate patient-specific machine learning models of glioblastoma invasion using transfer learning," *Amer. J. Neuroradiol.*, vol. 40, no. 3, pp. 418–425, 2019.

[33] A. Bellot and M. Schaar, "Boosting transfer learning with survival data from heterogeneous domains," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 57–65.

[34] S. Kiranyaz, T. Ince, R. Hamila, and M. Gabbouj, "Convolutional neural networks for patient-specific ecg classification," in *Proc. IEEE 37th Annu. Int. Conf. Eng. Med. Biol. Soc.*, 2015, pp. 2608–2611.

[35] P. G. Gibson, "Monitoring the patient with asthma: An evidence-based approach," *J. Allergy Clin. Immunol.*, vol. 106, no. 1, pp. 17–26, 2000.

[36] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 4107–4115.

[37] A. G. Howard *et al.*, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[38] T. Sheng, C. Feng, S. Zhuo, X. Zhang, L. Shen, and M. Aleksic, "A quantization-friendly separable convolution for mobilenets," in *Proc. 1st Workshop Energy Efficient Mach. Learn. Cogn. Comput. Embedded Appl.*, 2018, pp. 14–18.

[39] B. Rocha *et al.*, "A respiratory sound database for the development of automated classification," in *Prec. Med. Powered by pHealth and Con. Health*. Berlin, Germany: Springer, 2018, pp. 33–37.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representations*, San Diego, CA, USA, May 7–9, 2015.

[41] K. Kochetov, E. Putin, M. Balashov, A. Filchenkov, and A. Shalyto, "Noise masking recurrent neural network for respiratory sound classification," in *Proc. Int. Conf. Artif. Neural Netw.*, 2018, pp. 208–217.

[42] D. Perna, "Convolutional neural networks learning from respiratory data," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2018, pp. 2109–2113.

[43] G. Chambres, P. Hanna, and M. Desainte-Catherine, "Automatic detection of patient with respiratory diseases using lung sound analysis," in *Proc. Int. Conf. Content-Based Multimedia Indexing*, 2018, pp. 1–6.

[44] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 6, pp. 1291–1303, Jun. 2017.

[45] J. Sang, S. Park, and J. Lee, "Convolutional recurrent neural networks for urban sound classification using raw waveforms," in *Proc. 26th Eur. Signal Process. Conf.*, 2018, pp. 2444–2448.

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[47] P. Warden, "Speech commands: A public dataset for single-word speech recognition," vol. 1, 2017. [Online]. Available: http://download.tensorflow.org/data/speech_commands_v0

[48] R. X. A. Pramono, S. A. Imtiaz, and E. Rodriguez-Villegas, "Evaluation of features for classification of wheezes and normal respiratory sounds," *PloS One*, vol. 14, no. 3, 2019, Art. no. e0213659.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[50] S. Hershey *et al.*, "CNN architectures for large-scale audio classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 131–135.

[51] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 1041–1044.

[52] S. Alyamkin *et al.*, "Low-power computer vision: Status, challenges, opportunities," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 2, pp. 411–421, Jun. 2019.

[53] T. Gokmen, M. Rasch, and W. Haensch, "Training LSTM networks with resistive cross-point devices," *Front. Neurosci.*, vol. 12, 2018, Art. no. 745.

[54] D. Katz, "Fundamentals of embedded audio, part 3," Sep. 2007. [Online]. Available: https://www.eetimes.com/fundamentals-of-embedded-audio-part-3/

[55] W. Q. Lindh, M. Pooler, C. D. Tamparo, B. M. Dahl, and J. Morris, *Delmar's Comprehensive Medical Assisting: Administrative and Clinical Competencies*. Boston, Massachusetts, USA: Cengage Learning, 2013.

[56] A. Ignatov *et al.*, "AI benchmark: Running deep neural networks on android smartphones," in *Proc. Comput. Vision - ECCV 2018 Workshops*, Munich, Germany, vol. 11133, Sep. 2018, pp. 288–314, doi: 10.1007/978-3-030-11021-5_19.

[57] A. Basu *et al.*, "Low-power, adaptive neuromorphic systems: Recent progress and future directions," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 8, no. 1, pp. 6–27, Mar. 2018.

[58] J. Acharya, A. Patil, X. Li, Y. Chen, S. C. Liu, and A. Basu, "A comparison of low-complexity real-time feature extraction for neuromorphic speech recognition," *Front. Neurosci.*, vol. 12, p. 160, 2018, Art. no. 160.

Arindam Basu (Senior Member, IEEE) received the B.Tech. and M.Tech. degrees in electronics and electrical communication engineering from IIT Kharagpur, Kharagpur, India, in 2005, the M.S. degree in mathematics and the Ph.D. degree in electrical engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2009 and 2010, respectively. He joined Nanyang Technological University, Singapore, in 2010, where he is currently a tenured Associate Professor. His research interests include bio-inspired neuromorphic circuits, non-linear dynamics in neural systems, low-power analog IC design, and programmable circuits and devices. Dr. Basu was the recipient of the Best Student Paper Award from the ultrasonics symposium in 2006, best live demonstration at ISCAS 2010 and a finalist position in the Best Student Paper Contest at ISCAS 2008. He was also the recipient of the MIT Technology Reviews inaugural TR35@Singapore Award in 2012 for being among the top 12 innovators under the age of 35 in Southeast Asia, Australia, and New Zealand. He was the recipient of the Prime Minister of India Gold Medal in 2005 from IIT Kharagpur. He was a Guest Editor for two special issues in the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS for selected articles from ISCAS 2015 and BioCAS 2015. He is currently a Corresponding Guest Editor for the special issue on low power, adaptive neuromorphic systems: devices, circuit, architectures, and algorithms in the IEEE JOURNAL ON EMERGING TOPICS IN CIRCUITS AND SYSTEMS. He is currently an Associate Editor for the IEEE SENSORS JOURNAL, IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS, and *Frontiers in Neuroscience*. He was a Distinguished Lecturer of the IEEE Circuits and Systems Society from 2016 to 2017.

Jyotibdha Acharya (Student Member, IEEE) received the B.E. degree from the Department of ECE, Jadavpur University, Kolkata, India, in 2014. He is currently working toward the Ph.D. degree in HealthTech NTU, Interdisciplinary Graduate Program with Nanyang Technological University, Singapore. He worked in Philips Healthcare as a Graduate Engineer Trainee (Imaging Systems). His research interests include deep learning, neuromorphic algorithms, and biomedical applications.