Multimedia Tools and Applications (2023) 82:36615–36631
https://doi.org/10.1007/s11042-023-14727-0

A lung sound recognition model to diagnoses the respiratory diseases by using transfer learning

Kumari Nidhi Lal[1]

Received: 29 April 2022 / Revised: 29 September 2022 / Accepted: 5 February 2023 /
Published online: 29 March 2023

Abstract
Respiratory disease is one of the leading causes of death in the world. Through advances in Artificial Intelligence, it appears possible for the days of misdiagnosis and treatment of respiratory disease symptoms rather than their root cause to move behind us. The traditional convolutional neural network cannot extract the temporal features of lung sounds. To solve the problem, a lung sounds recognition algorithm based on VGGish- stacked BiGRU is proposed which combines the VGGish network with the stacked bidirectional gated recurrent unit neural network. A lung Sound Recognition Algorithm Based on VGGish-Stacked BiGRU is used as a feature extractor which is a pre-trained model used for transfer learning. The target model is built with the same structure as the source model which is the VGGish model and parameter transfer is done from the source model to the target model. The multi-layer BiGRU stack is used to enhance the feature value and retain the model. While fine-tuning of the parameter of VGGish is frozen which successfully improves the model. The experimental results show that the proposed algorithm improves the recognition accuracy of lung sounds and the recognition accuracy of respiratory diseases.

Keywords Internet of things · Energy · Wireless sensor network · Performance analysis

1 Introduction

Breath sounds are the noises produced by the structures of the lungs during breathing. The lung sounds are best heard with a stethoscope. Using a stethoscope, the doctor may hear normal breathing sounds, decreased or absent breath sounds, and abnormal breath sounds. Auscultation serves as a quick and inexpensive way for the modern-day physician to infer a variety of disease states about the cardiovascular, respiratory, and gastrointestinal systems, which allows for streamlined diagnoses and management [9]. To optimize the effectiveness of auscultation the surroundings should be Quiet, Warm, and with Appropriate lighting. Therefore, auscultation results are affected by the external environment as well

✉ Kumari Nidhi Lal
nidhilal@cse.vnit.ac.in

[1] Department of Computer Science Engineering, Visvesvaraya National Institute of Technology (VNIT Nagpur), Nagpur, Maharashrta, India

as the doctor's medical experience and hearing condition. Acoustic stethoscopes operate on the transmission of sound from the chest piece, via air-filled hollow tubes, to the listener's ears. The problem with acoustic stethoscopes is that the sound level is extremely low [10].

In recent years, an electronic stethoscope (or stethoscope) overcomes low sound levels by electronically amplifying body sounds. However, amplification of stethoscope contact artifacts and component cutoffs (frequency response thresholds of electronic stethoscope microphones, pre-amps, amps, and speakers) limit electronically amplified stethoscopes' overall utility by amplifying mid-range sounds, while simultaneously attenuating high- and low- frequency range sounds. Currently, several companies offer electronic stethoscopes. Electronic stethoscopes require the conversion of acoustic sound waves to electrical signals which can then be amplified and processed for optimal listening. Unlike acoustic stethoscopes, which are all based on the same physics, transducers in electronic stethoscopes vary widely. Electronic stethoscopes are also used with computer-aided auscultation programs to analyze the recorded heart sounds pathological or innocent heart murmurs [12]. Computerized recorded lung sounds can be analyzed using time series and may offer an approach to the diagnosis via the recognition model. Moreover, it can predict lung diseases like asthma, Chronic Obstructive Pulmonary Disorder (COPD), and health status. Features are extracted using a discrete wavelet transform and a Decision Tree Classifier is used for early predictions on symptom-based COPD exacerbations using Artificial Intelligence. Wavelet decomposition is applied to each t–f image representation of the EEG signals resulting in Diagonal (D), Vertical (V), and Horizontal (H) components which are stored as images and are employed for feature extraction in [22]. It could effectively recognize the polyphonic lung sounds and the sharp lung sounds. Mel Frequency Cepstral Coefficients (MFCC) are extracted from the pre-processed pulmonary acoustic signals.

The performance of Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) classifiers in the diagnosis of respiratory pathologies have been compared by using respiratory sounds from the R.A.L.E database in [17]. Lung sound signals are decomposed into the frequency subbands using wavelet transform and a set of statistical features are extracted from the subbands to represent the distribution of wavelet coefficients. Lung sounds are classified into six categories by using an artificial neural network in [14]. The lung sounds are classified by the Back Propagation (BP) neural network. The wavelet packet decomposition is used to get the energy of the lung sounds with different frequency ranges, which are taken as features to recognize four kinds of lung sounds including normal, tracheitis, pneumonia, and asthma [2]. To identify chronic obstructive pulmonary disease using applied machine learning algorithms and forced oscillation measurements, they Worked on selecting features for various classification algorithms such as linear bayesian, KNN, decision trees, Artificial Neural Network (ANN), and SVM in [6]. In this study for the Time–Frequency (TF) analyses, 64-point Wave-Function Theory (WFT) with Gaussian, Blackman, Hanning, Hamming, Bartlett, Triangular, and Rectangular windows are used. In WFT, shifting the analysis window by less than the window length results in an overlapped Fast Fourier Transform (FFT) of the analyzed signals in [27].

Consequently, to relieve patients of the inconvenience caused by their symptoms, new methods must be developed for identifying respiratory diseases [25]. Incorporating computer analysis and electronic auscultation in the study of lung sounds ensures high accuracy and timely diagnosis [4]. It eliminates the subjectivity of the listener and identifies pathological features that physicians cannot identify. With computer analysis, physicians can provide accurate diagnoses and initiate treatment early, relieving the discomfort of their patients [11]. Currently, several studies have attempted to incorporate computer algorithms using

Convolutional Neural Networks (CNNs) for detecting adventitious lung sounds [3]. However, most of these works focused only on lung sound classification instead of segmentation of the sound signal [20]. The main contributions of the paper are:

- The splitting of respiratory cycles into phases is done and perform sample padding on both of them to enrich the information of adventitious sounds for the lung sound classification system.
- Utilized transfer learning model in using the pre-trained single input model to build a multi-input VGGish model for lung sound classification. The outline of the paper is as follows: In Section 2, a literature survey is introduced that describes the lung sound databases used as source and target domains. In Section 3, the proposed work is presented. In Section 4, experimental setup including the evaluation metrics and the experimental results are described. Finally, the conclusion of the work is done in Section 5.

2 Related work

The related work section describes the stage of lung sound recognition using traditional machine learning methods. With the development of deep learning in recent years, lung sound diagnosis technology is improved and new methods are developed. MFCCs are coefficients that collectively make up a Mel-Frequency Cepstrum (MFC). An MFC is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrogram on a non-linear model scale of frequency. These features represent phonemes (the distinct units of sound) as the shape of the vocal tract (which is responsible for sound generation) is manifest [11]. This makes MFCC a great feature to consider for respiratory audio analysis. Adam is a combination of stochastic gradient descent and RMSprop algorithm, which provides an improved network weight optimization logic that also makes the process of hyperparameter tuning more efficient [3, 7].

They employed two types of machine learning algorithms; MFCC features in a Support Vector Machine (SVM) and spectrogram images in the Convolutional Neural Network (CNN). Since using MFCC features with an SVM algorithm is a generally accepted classification method for audio, the results can be utilized to benchmark the CNN algorithm. The four data sets are prepared for each CNN and SVM algorithm to classify respiratory audio in [8]. The method extracted MFCC features, and a 2-Layer CNN (2L-CNN) is used to train and recognize the lung sounds. Experimental results showed that the recognition method based on CNN is before the method based on SVM. MFCC features are utilized to identify the lung sounds by a 5-Layer CNN (5L- CNN), and better recognition results are obtained in [1]. They used a variant of 3D VoxResNet for COPD emphysema classification. The model uses volume-wise annotations without any further feature enhancement or addition of metadata. For the emphysema classification, they fine-tuned the COPD classification network, which significantly increased the model performance in [18]. A combined model framework DNN-HMM is proposed to identify the normal and abnormal lung sounds, which combined a deep neural network with the hidden Markov model in [29]. VGGish is firstly used to overcome the dependence of the algorithms on data and features [25]. Secondly, the temporal feature of the lung sound signals is captured by taking the bidirectional gated recurrent unit neural network (BiGRU) as the retraining layer of transfer learning [4, 26]. In [21] Data augmentation in both the time domain and time-frequency domain is used to account for

the class imbalance of the ICBHI and our multi-channel lung sound dataset. Additionally, the author introduced a spectrum correction to account for the variations of the recording device properties on the ICBHI dataset. In [19], authors combined well-known strategies for pre-processing, feature extraction, and classification which brings us to a remarkable success rate of disease and its severity recognition with an accuracy of 96.05% (97.7% for Non-Severe COVID-19 images and 93% for Severe COVID-19 images). The author mentioned that their model can therefore help radiologists detect COVID-19 and the extent of its severity. In [13], an extensive COVID-19 X-ray and CT Chest Images Dataset has been used and a generative adversarial network (GAN) coupled with trained, semi-supervised CycleGAN (SSA- CycleGAN) has been applied to augment the training dataset. Then a newly designed and finetuned Inception V3 transfer learning model has been developed to train the algorithm for detecting the COVID-19 pandemic. In this paper, a lung sounds recognition algorithm based on deep learning and transfer learning. VGGish is firstly used to overcome the dependence of the algorithms on data and features and then by taking the stacked multilayer bidirectional gated recurrent unit neural network as a retaining layer of transfer learning, the temporal feature is captured.

3 Proposed work

The idea of transfer learning is used for utilizing knowledge acquired for one task to solve related ones and is commonly used for deep learning. In the model, the pre-trained VGGish model is used as a starting point to learn a new task. From the famous image recognition network VGG, TensorFlow Model garden has developed a research model associated with AudioSet.Due to the similarities in architecture, this sound classifier is given the name VGGish. More specifically, it is adapted from VGG16 Configuration along with some minor changes. In the model, the source domain data is google Audio set, and target data is gathered from the respiratory sound database from Kaggle.

VGGish network part is a transfer from the source domain to the target domain and the parameter of the VGGish network is loaded into the network. The output of the network is the 128- dimensional feature vector for each time-series information. Then multilayer bidirectional gated recurrent unit neural network is used for processing better time-series data and to achieve even greater results, stack multiple Recurrent Neural networks (RNN)(Long short-term memory (LSTM) or Gated Recurrent Units (GRU) or normal RNN) can be stacked on top of each other. Here, the stacked bidirectional gated recurrent unit neural network is taken as the retraining layer in the model, and the problem of insufficient data is compensated by freezing the parameters of the network layer of the pretraining model. The first bidirectional GRU network part inputs the features extracted by the transfer layer into the BiGRU network. The first bidirectional gated recurrent unit neural network output sequence rather than a single value output to the layer below. It allows more complex input patterns can be described at every layer. Since the similarity between the target data and the source domain data is lower, it is very indispensable to retrain the model with the target data. To this end, the BiGRU network is taken as the retraining layer in the model, and the problem of insufficient data is compensated by freezing the parameters of the network layer of the pretraining model as shown in Fig. 1. RNN is chosen as the retraining part of the model since RNN has a strong ability to capture the time-series features of the signals, and can pay more attention to the data context. As a kind of time-series data, the temporal relationship existing in the lung sounds can effectively be captured through retraining BiGRU.

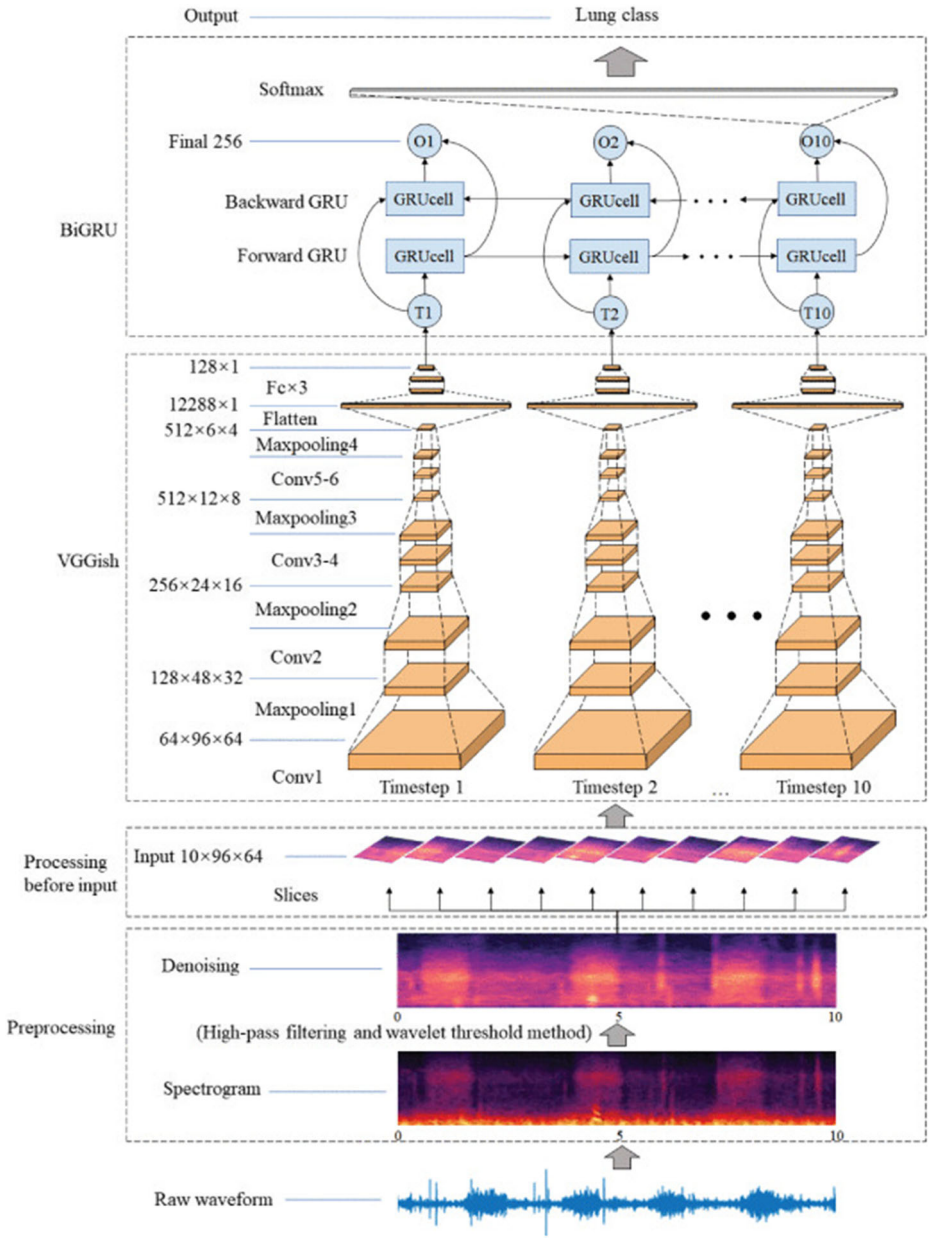

Fig. 1 The proposed model architecture

3.1 Data preprocessing

Nonlinear time series denoising is the prerequisite for extracting effective information from observation sequences. An effective chaotic signal denoising method not only has a good

Signal-to-Noise Ratio (SNR) enhancement performance but also can remain as a good unpredictable denoised signal. Denoising can be removed by a high-a-pass filter or low-pass filter. A High-Pass Filter (HPF) is an electronic filter that passes signals with a frequency higher than a certain cutoff frequency and attenuates signals with frequencies lower than the cutoff frequency. The amount of attenuation for each frequency depends on the filter design [28]. The complexity of filter type is defined by the filter's "order", which is dependent upon the number of reactive components such as capacitors or inductors within its design. The rate of roll-off and the width of the transition band depends upon the order number of the filter and a simple first-order filter has a standard roll-off rate of 20dB/decade or 6dB/octave as shown in Fig. 2.

Then, for a filter that has an nth number order, it will have a subsequent roll-off rate of 20n dB/decade or 6n dB/octave. Therefore, a first-order filter has a roll-off rate of 20dB/decade (6dB/octave), a second-order filter has a roll-off rate of 40dB/decade (12dB/octave), and a fourth-order filter has a roll-off rate of 80dB/decade (24dB/octave), etc, etc.High-order filters, such as third, fourth, and fifth-order are usually formed by cascading together single first-order and second-order filters. The low-frequency noise under100Hz can be removed by high-pass filters as the frequency band of the lung sound signals is 100HZ to 2000Hz. The cutoff frequency is 100Hz. The cut-off frequency is normalized by the sampling frequency. The frequency band of the heart sounds in the lung sounds is 5HZ to 600Hz, which highly coincides with the low-frequency part of the lung sounds as shown in Fig. 3. It is difficult to remove the interference of the heart sounds without damaging the lung sounds by simple filtering.

Wavelets have been a powerful tool to decompose the audio signal into parts and apply thresholds to eliminate unwanted signal-like noise. The thresholding method is the most important in the process of Audio De nosing. The wavelet threshold denoising method is proposed by American scholar Donohue. The method is simple to calculate and the noise can be suppressed to a large extent. At the same time, singular information of the original signal can be preserved well. Therefore, it is a simple and effective method. The thresholding used is the VisuShrink method or the universal threshold introduced by Donoho. The VisuShrink approach employs a single, universal threshold for all wavelet detail coefficients. This threshold is designed to remove additive Gaussian noise with high probability, which tends to result in an overly smooth image appearance as shown in Fig. 4. By

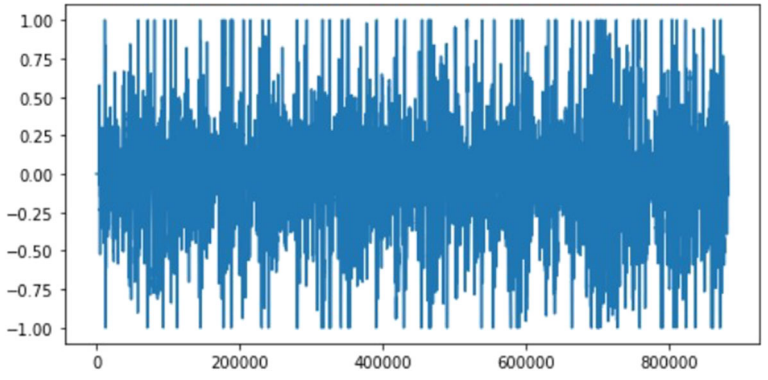

Fig. 2 Input waveform before data processing

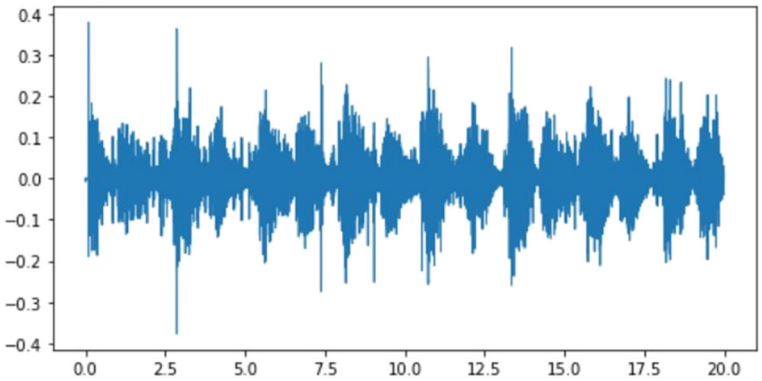

Fig. 3 After denoising after data processing

specifying a sigma that is smaller than the true noise standard deviation, a more visually agreeable result can be obtained. Among these methods, the universal threshold is the most widely used because of its simpleness and effectiveness. The formula for the universal threshold is expressed as follows: where is the average variance of the noise and the signal length is calculated using the median estimate method.

3.2 Input processing of the lung sound data

The audio samples are usually represented as time series, where the y-axis measurement is the amplitude of the waveform. The amplitude is usually measured as a function of the change in pressure around the microphone or receiver device that originally picked up the

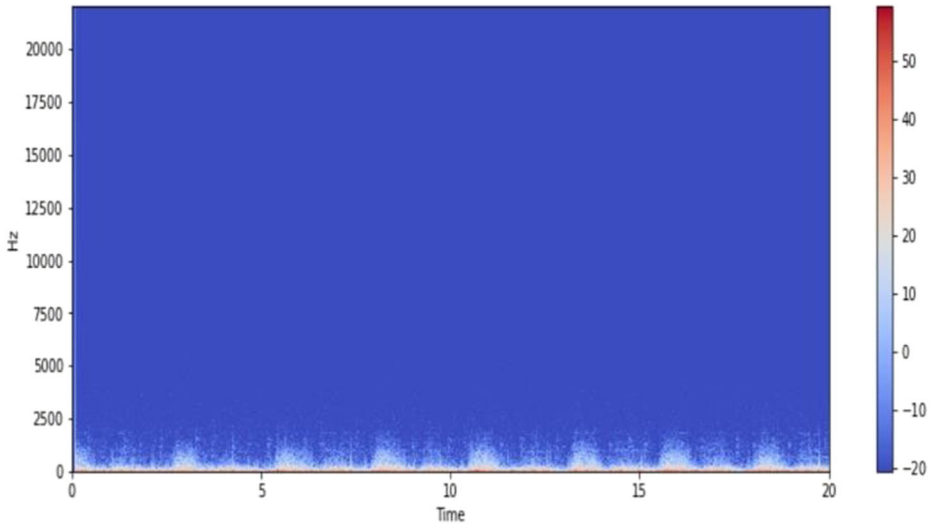

Fig. 4 Spectrogram of the input wave

audio. The model takes Mel spectrogram features of the lung sounds as the input of the network. To process the Mel spectrogram the following steps need to be done. The source domain dataset is trained on millions of seconds of audio samples collected from YouTube videos, resampled to 16 kHz mono. Therefore, the target data is resampled to 16 kHz mono. The noise suppression ability of FM decreases with the increase in the frequencies.

Thus increasing the relative strength or amplitude of the high-frequency components of the message signal before modulation is termed Pre-emphasis. Therefore, pre-emphasis is done after resampling. The audio signal is comprised of several single-frequency sound waves. When taking samples of the signal over time, capturing the resultant amplitudes will be done. By mapping the audio signal from the time domain to the frequency domain using the fast Fourier transform. The Fourier transform is a mathematical formula that allows us to decompose a signal into its frequencies and the frequency's amplitude. In other words, it converts the signal from the time domain into the frequency domain. The result is called a spectrum as shown in Fig. 5.

The fast Fourier transform (FFT) is an algorithm that can efficiently compute the Fourier transform [24]. It is widely used in signal processing. The FFT is computed on overlapping windowed segments of the signal, and hence it is called the spectrogram. The lung sounds are transformed from the time domain to the frequency domain by performing STFT. In the time-frequency transformation, it is necessary to define a range of the frequency domain of the lung sounds to obtain the main frequency domain information. The sampling frequency of the lung sounds is 44100Hz. Gathering a local Fourier transform at an equispaced point creates a local Fourier transform, also called a spectrogram. By carefully choosing the window, this transform corresponds to the decomposition of the signal in a redundant tight frame. The redundancy corresponds to the overlap of the windows, and the tight frame corresponds to the fact that the pseudo-inverse is simply the transposed of the transform (it means that the same window can be used for synthesis with a simple summation of the reconstructed signal over each window). The only parameters of the transform are the size of the window and the overlap. They used STFT with a periodic Hann window, a window size of 25 ms, and a window hop of 10 ms to generate a spectrogram for each sample. A spectrogram is a visual depiction of a signal's frequency composition over time. the

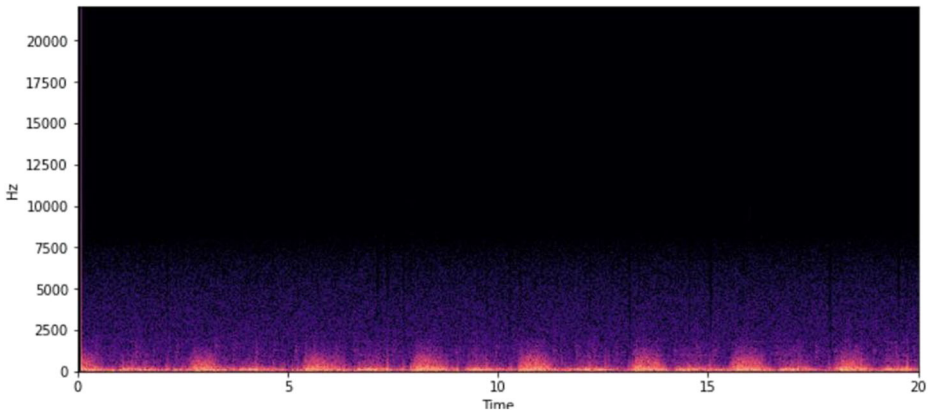

Fig. 5 Spectrogram of input after denoising

spectrum line energy of the lung sounds is filtered by using a Mel filter bank. The function of filters is expressed using (1).

$$H_m(k) = \begin{cases} 0 & k < f(m-1) \\ \frac{k-f(m-1)}{f(m)-f(m-1)} & f(m-1) \le k \le f(m) \\ \frac{f(m+1)-k}{f(m+1)-f(m)} & f(m) < k \le f(m+1) \\ 0 & k > f(m+1) \end{cases} \tag{1}$$

where $0 \geq m \geq M$, and M is the number of filters. Its center frequency can be expressed using (2).

$$f(m) = \left(\frac{N}{f_s}\right) F_{mel}^{-1}(F_{mel}(f_l) + m\left(\frac{F_{mel}(f_h) - F_{mel}(f_l}{M+1}\right) \tag{2}$$

The Mel spectrogram is used to provide our model with sound information similar to what a human would perceive. A mel spectrogram is a spectrogram where the frequencies are converted to the mel scale. The mel scale (after the word melody) is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The reference point between this scale and normal frequency measurement is defined by assigning a perceptual pitch of 1000 mels to a 1000 Hz tone, 40 dB above the listener's threshold. The Mel scale provides a linear scale for the human auditory system and is related to Hertz by the following formula, where m represents Mels and f represents Hertz using (3).

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{3}$$

The amplitude spectrum obtained by the short-time Fourier transform is separately multiplied with each filter and all items are accumulated. Finally, Mel spectrogram is achieved by taking the log value of the energy and expanding it in the time domain. Each frame contains 64 Mel bands while extracting Mel spectrogram features. Every one of these spectrograms is mapped into 64 mel bins covering the range 125-7500 Hz, resulting in so-called mel spectrograms. Then, the creation of stabilized log mel spectrograms is done by taking the log of (mel-spectrum + 0.01), which is framed into non-overlapping examples of 0.96 s. Every example covers 64 mel bands and 96 frames of 10 ms each.

3.3 Transferring the model parameters

From the Source Domain to the Target Domain, a transfer-learning problem is considered by using the parameter transfer approach, where a suitable parameter of feature mapping is learned through one task and applied to another objective task. The transferring model is shown in Fig. 6. VGGish network in the dotted box on the right needs to be transferred to the target model the structure of the source and the target must be the same for transferring parameters. Therefore, the VGGish network with the same structure is firstly built on the target network and parameters are loaded by loading the downloaded file from [31] Where the tensor node of the checkpoint in TensorFlow is converted into .h5 format. Thus, the parameters of the corresponding layers in the source domain network are transferred to the target network [23]. Deep learning systems and models are layered architectures that learn different features at different layers (hierarchical representations of layered features). These layers are then connected to the last layer (usually a fully connected layer, in the case

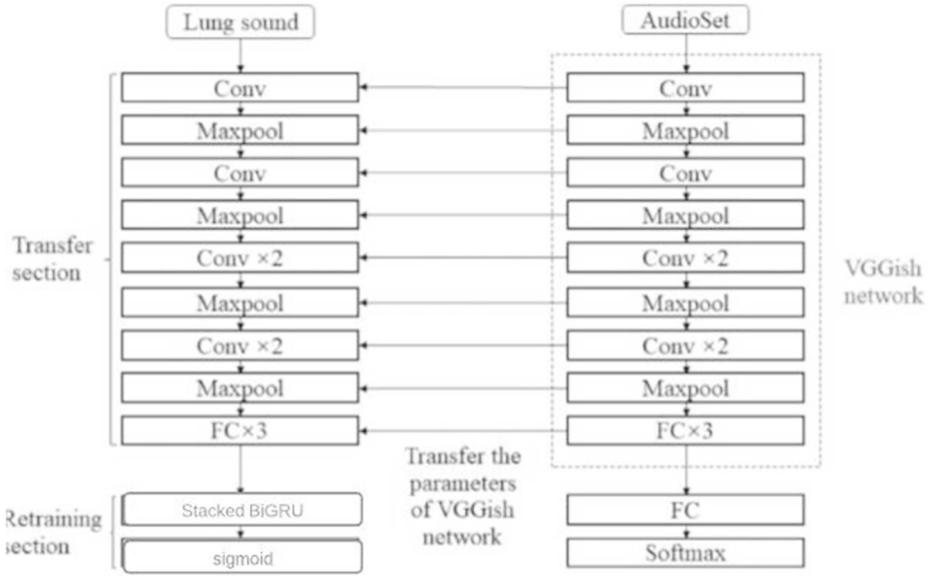

Fig. 6 Transfer learning model

of supervised learning) to get the final output. The layered architecture utilizes a pre-trained network without its last layer as a fixed feature extractor for other tasks [15]. Therefore in this model, the last collected layer is removed and a stacked bidirectional layer is used. During fine-tuning, the source model parameter is frozen because the insufficient target data can be compensated by freezing the network parameters of the pre-trained model as well as not fixing weight will affect the basic knowledge learned by the model when backpropagated and feature representation is also damaged while retaining which is need to learn new objects and labels [30].

Fine-tuning, in general, means making small adjustments to a process to achieve the desired output or performance. Fine-tuning deep learning involves using weights of a previous deep learning algorithm for programming another similar deep learning process [5]. Here the source model parameter is first transferred using parameter transfer then the model is frozen and the remaining layers are passed through backpropagation where weights are changed for the remaining layers while fine-tuning. Therefore, Unfreeze a few of the layers of a frozen model base and jointly train both the newly-added classifier layers and the last layers of the base model. This allows us to "fine-tune" the higher-order feature representations in the base model to make them more relevant for the specific task.

3.4 Retraining of BiGRU network

Recurrent Neural Networks (RNN) are designed to work with sequential data because standard feedforward neural networks cannot handle speech data well (due to lacking a way to feed information from a later layer back to an earlier layer). RNNs have been introduced to take the temporal dependencies of speech data into account. Furthermore, RNNs cannot handle the long-term dependencies due to the vanishing/exploding gradient problems in an accurate manner [16]. Therefore, LSTMs and a few years later GRUs were introduced to

overcome the shortcomings of RNNs. The workflow of GRU is the same as RNN. However, the difference is in the operations inside the GRU unit as shown in Fig. 7.

Two gates are reset and update two gates are reset update. These gates are neural networks, each gate has its weights and biases. Fewer parameters mean GRUs are generally easier/faster to train than their LSTM counterparts. The update gate decides whether the cell state should be updated with the candidate state(current activation value)or not. The reset gate is used to decide whether the previous cell state is important or not. In a few scenarios, the reset gate is not used in simple GRU. It is just the same as the hidden state(activation) of RNN. The last cell state is dependent on the update gate. It may or may not be updated with the candidate's state. In GRU the final cell state is directly passed as the activation to the next cell. The specific calculation process :

$$\begin{cases} Z_t = \sigma\left(W_i * \left[h_{t-1,x_t}\right]\right) \\ r_t = \sigma\left(W_r * \left[h_{t-1,x_t}\right]\right) \\ h_t = \tanh\left(W_c * \left[r_t * h_{t-1}, x_t\right]\right) \\ Z_t = (1 - z_t) * c_{t-1} + z_t * h_t \end{cases} \tag{4}$$

A Bidirectional GRU, or BiGRU, is a sequence processing model that consists of two GRUs. One takes the input in a forward direction, and the other in a backward direction. It is a bidirectional recurrent neural network with only the input and forgets gates. In the forward propagation of the network, the VGGish network will output a 10×128 feature vector for each 10-second lung sound, and randomly initializes a bidirectional GRU network. The 10×128 feature vector will be input into the BiGRU network in time series. The output of the layer is then passed to the next layer as input which contains 10 ×256 vectors for each 10-seconds. Therefore, the second bidirectional gated recurrent unit layer output 256 vector is then given to the dense layer which has an activation layer as sigmoid. Then, dense layer output is used to predict the status of health, and Chronic obstructive pulmonary disease (COPD) is a chronic inflammatory lung disease. Here multi-layer BiGRU stack is used to

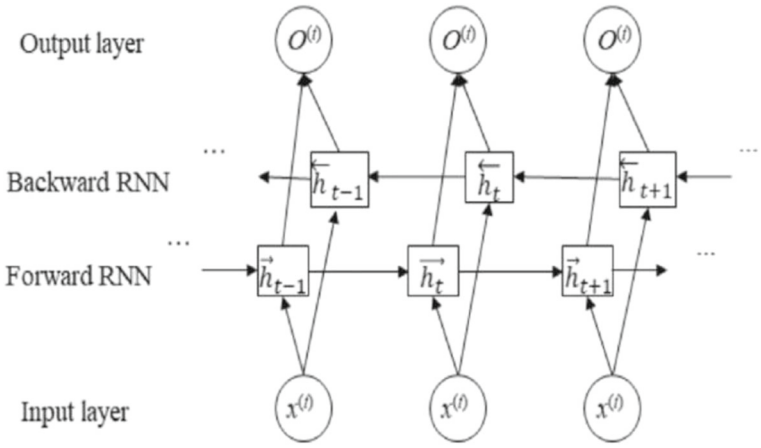

Fig. 7 Structure of GRU

enhance the feature value to obtain more long-distance context information while avoiding the gradient disappearance and gradient explosion problems caused by ordinary recurrent neural networks. During the backpropagation, the learning rate is slow and in the VGGish-stacked BiGRU network, only stacked BiGRU layers are fine-tuned since the parameters of VGGish layers are frozen. For the classification problem, the sigmoid loss is taken as the loss of the function of the backpropagation. The output of the sigmoid layer is computed by using (4).

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

Cross-entropy loss, or log loss, measures the performance of a classification model whose output is a probability value between 0 and 1. Cross-entropy loss increases as the predicted probability diverge from the actual label using (5). Therefore, predicting a probability of .012 when the actual observation label is 1 would be bad and result in a high loss value. A perfect model would have a log loss of 0.

4 Experimental results

4.1 Experimental data

The Respiratory Sound database is originally compiled to support the scientific challenge organized at Int. Conf. on Biomedical Health Informatics - ICBHI 2017. The current version of this database is made freely available for research and contains both the public and private datasets of the ICBHI challenge. The database consists of a total of 5.5 hours of recordings containing 6898 respiratory cycles, of which 1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes, in 920 annotated audio samples from 126 subjects. 36 each healthy and Chronic obstructive pulmonary disease (COPD) is a chronic inflammatory lung disease audio file is used in the experiment. Used validation split argument of the fit method to use 15 % of your train data as validation data (Tables 1, 2 and 3).

4.2 Parameter settings

For the learning of the network, the setting of parameters is particularly important, it will directly affect the quality of the network. The specific settings of the model parameters are shown below:

Table 1 Parameter setting for proposed working model

S.No.	Parameter	value in simulation
1	Learning rate	0.0001
2	Batch size	16
3	Epochs	50
4	beta_1	0.1
5	beta_2	0.999
6	Optimizer	Adam

Table 2 Confusion matrix of the previous model

S.No.	Precision	Recall	f1-score	Support
Healthy	0.85	1.00	0.92	70
Copd	1.00	0.83	0.91	70
Accuracy			0.91	140
Macro avg	0.93	0.91	0.91	140
Weighted avg	0.93	0.91	0.91	140

4.3 Evaluation index

The evaluation indicators in this article are precision, recall, and f1-score, support. Precision and recall are two numbers that together are used to evaluate the performance of classification or information retrieval systems. Precision is defined as the fraction of relevant instances among all retrieved instances. Recall sometimes referred to as 'sensitivity, is the fraction of retrieved instances among all relevant instances. A perfect classifier has precision and recalls both equal to 1. It is often possible to calibrate the number of results returned by a model and improve precision at the expense of a recall using (6), or vice versa. Precision and recall should always be reported together. Precision and recall are sometimes combined into the F-score if a single numerical measurement of a system's performance is required using (7).

$$precision = \frac{True\ Positive(t_p)}{True\ Positive(t_p) + False\ Positive(f_p)} \tag{6}$$

$$recall = \frac{True\ Positive(t_p)}{True\ Positive(t_p) + False\ Negative(f_n)} \tag{7}$$

If a single number is required to describe the performance of a model, the most convenient figure is the F-score, which is the harmonic mean of the precision and recall using (8):

$$F = 2 * \frac{precision * recall}{precision + recall} \tag{8}$$

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{9}$$

Accuracy is defined as the percentage of correct predictions for the test data. It can be calculated easily by dividing the number of correct predictions by the number of total predictions. The sum of the scores of all classes after multiplying their respective class

Table 3 Confusion matrix of proposed model

S.No.	Precision	Recall	f1-score	Support
Healthy	0.90	1.00	0.95	70
Copd	1.00	0.89	0.94	70
Accuracy			0.94	140
Macro avg	0.95	0.94	0.94	140
Weighted avg	0.95	0.94	0.94	140

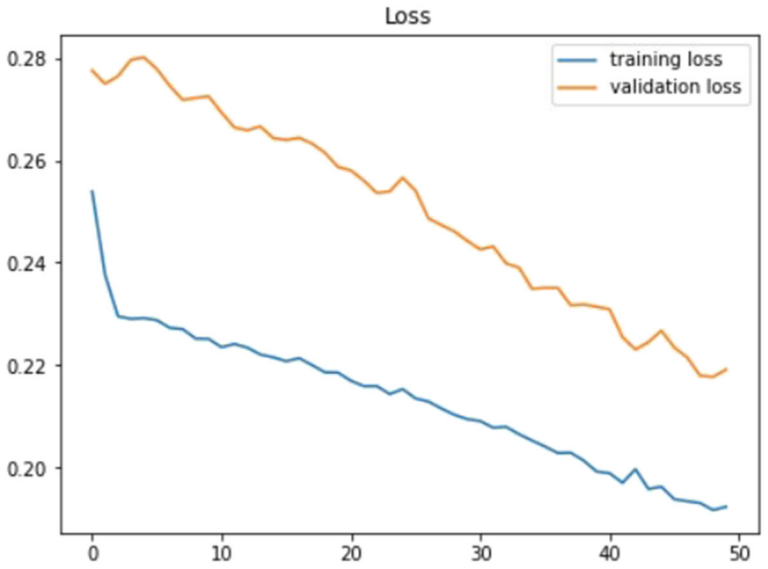

Fig. 8 Training and validation loss curve

proportions is Weighted average scores. It is the simple mean of scores of all classes is Macro-average scores.

$$[[\,70\;\;0\,][\,12\;\;58\,]]$$

4.4 Result analysis

The confusion matrix of the model with fine-tuning and single layer Bidirectional Gated Recurrent Unit is shown in Table II. The confusion matrix of the model with fine-tuning and stacked Bidirectional Gated Recurrent Unit is shown in Table III. To validate the effect of fine-tuning the comparison between the confusion matrix is used. The result showed improvement in the model due to fine-tuning the stacked Bidirectional Gated Recurrent Unit and little signs of overfitting on the data. Without fine-tuning the model the healthy and COPD precision of the model is around 92 and 97 percent respectively. However, after fine-tuning the model rose to 90 and 1 percent. During transferring the model parameters, although the transfer learning can bring the common features learned from the source domain data, there are only a few cardiopulmonary sound data in the source domain data. In addition, there is less similarity between the target data and the source domain data. Therefore, it is necessary to retrain the model for extracting some unique features from the target dataset. It will better improve the recognition accuracy. The model training and validation loss curve is shown in Fig. 8.

$$[[\,70\;\;0\,][\,8\;\;62\,]]$$

5 Conclusion

This paper proposes a lung Sound Recognition Algorithm Based on VGGish-Stacked BiGRU.VGGish model is used as a feature extractor which is a pre-trained model used for

transfer learning. The target model is built with the same structure as the source model which is the VGGish model and parameter transfer is done from the source model to the target model. The multi-layer BiGRU stack is used to enhance the feature value and retain the model. while fine-tuning the model the parameter of VGGish is frozen which successfully improves the model. Finally, it is concluded that the best classification effect can be obtained by introducing VGGish and two-layer bidirectional gated recurrent. But because the network is more complicated, the small dataset used affects the accuracy of the model and shows signs of overfitting on the data. In the future, it is the goal of the next step to studying how to improve classification accuracy and reduce overfitting. The audio set has less similarity with the source dataset which also affects the accuracy. In addition, the performance of the the proposed system can be considered to examine the clinically obtained CT scan slices with a lung infection. Further, the proposed methodology needs to be investigated on the a larger set of databases of CT scan images of the patients.

Funding I acknowledge that there is no external funding.

Data Availability The Respiratory Sound database was originally compiled to support the scientific challenge organized at Int. Conf. on Biomedical Health Informatics - ICBHI 2017. The current version of this database is made freely available for research and contains both the public and the private datasets of the ICBHI challenge. I can provide the dataset, if needed.

Declarations

Conflict of Interests I acknowledge that there is no external conflicts of interests.

Competing interests I acknowledge that there is no external competing interests

References

1. Ahmed J, Vesal S, Durlak F, Kaergel R, Ravikumar N, Rémy-jardin M, Maier A, Tolxdorff T, Deserno T, Handels H, Maier A (2020) COPD classification in CT images using a 3D convolutional neural network. In: Maier-Hein K, Palm C (eds) Bildverarbeitung für die medizin 2020—informatik aktuell. Wiesbaden, Springer Vieweg, pp 39-45, vol 2020
2. Amaral JLM, Lopes AJ, Jansen JM, Faria ACD, Melo PL (2011) Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease ,pubmed.ncbi.nlm.nih.gov
3. Ambati LS (2019) Human activity recognition : a comparison of machine learning approaches. J Midwest Assoc Inf Syst 2021(1):49
4. Ambati LS, El-Gayar O (2020) Influence of the digital divide and socio-economic factors on prevalence of diabetes. Issues Inf Syst 21(4):103–113
5. Anusha N, Kahn G, Fearing RS, Levine S (2018) Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, pp 7559-7566
6. Aydin N (2000) Optimization of processing parameters for the analysis and detection of embolic signals. Eur J Ultrasound, pp 79
7. Aykanat M, Kılıç Ö, Kurt B, Sanyal S (2017) Classification of lung sounds using convolutional neural networks. EURASIP J Image Video Proc 2017(1):65. https://doi.org/10.1186/s13640-017-0213-2
8. Bardou D, Zhang K, Ahmad SM (2018) Lung sounds classification using convolutional neural networks. Artif Intell Med 88:58–69
9. Caleb AN, Roda MD (2019) Modern-day cardiac auscultatory teaching and its role alongside echocardiography, BCMJ, pp 128-130
10. Dredge S (2009) Auscultation, physio-pedia.com, p

11. El-Gayar OF, Ambati LS, Nawar N (2020) Wearables, artificial intelligence, and the future of healthcare. In: AI and Big Data's Potential for Disruptive Innovation, IGI Global, (pp 104-129)
12. Fernandez-Granero MA (2018) An artificial intelligence approach to early predict symptom-based exacerbations of COPD, pp 778–784
13. Ghazal B, Zhou X, Barua PD, Gururajan R, Li Y, Acharya UR (2022) Application of CycleGAN and transfer learning techniques for automated detection of COVID-19 using X-ray images. Pattern Recog Lett 153:67–74
14. Haider NS (2020) Feature Extraction and Classification Methods for Lung Sounds. Int J Innov Technol Exploring Eng, pp 10, ISSN: 2278-3075 volume-10 Issue-1
15. Huang R, Hansen JH (2006) Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora. IEEE Trans Audio Speech Lang Process 14(3):907–919
16. Irum H, Ahmad M, Hussain A, Ashraf MU, Saeed IA, Qadri SF, Alghamdi AM, Alfakeeh AS (2021) Breast cancer classification from histopathological images using patch-based deep learning modeling. IEEE Access 9:24273-24287
17. Kandaswamy A, Sathish Kumar CSC, Ramanathan RP, Jayaraman S, Malmurugan N (2004) Neural classification of lung sounds using wavelet coefficients,pubmed.ncbi.nlm.nih.gov
18. Li L, Xu W, Hong Q, Tong F, Wu J (2017) Classification between normal and adventitious lung sounds using deep neural network. Proc 10th Int Symp Chin Spoken Lang Process (ISCSLP), pp 1–5
19. Mansi G, Swaraj A, Verma K (2022) Classification of COVID-19 patients with their severity level from chest ct scans using transfer learning. arXiv:2205.13774
20. Nguyen T, Pernkopf F (2021) Crackle detection in lung sounds using transfer learning and multi-input convolutional neural networks. In: 2021 43rd annual international conference of the ieee engineering in medicine & biology society (EMBC), IEEE, pp 80–83. https://doi.org/10.1109/EMBC46164.2021.9630577
21. Nguyen T, Pernkopf F (2022) Lung sound classification using co-tuning and stochastic normalization. IEEE Trans Biomed Eng 69(9):2872–2882. https://doi.org/10.1109/TBME.2022.3156293
22. Palaniappan R, Sundaraj K, Sundaraj S (2014) A comparative study of the SVM and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals biomedcentral.com. pp 15–223
23. Rekha M (2019) Churning the confusion out of the confusion matrix",blog.clairvoyantsoft.com
24. Ruder S (2016) An overview of gradient descent optimization algorithms. arXiv:1609.04747
25. Sai AL, Omar E-G, Nevine N (2021) Design principles for multiple sclerosis mobile self-management applications : a patient-centric perspective. AMCIS 2021 Proceedings 11
26. Sarkar DDJ (2018) A comprehensive hands-on Guide to Transfer Learning with Real-World Applications in Deep Learning. Towards Data Science
27. Sharma N, Sharma R, Jindalb N (2021) Machine learning and deep learning applications-a vision. Glob Transit Proc 2(1):24–28
28. Shewalkar A, Nyavanandi D, Simone A (2019) Performance evaluation of deep neural networks applied to speech recognition : RNN, LSTM, and GRU. J Artif Intell Soft Comput Res, pp 235–245
29. Shi L, Kang DU, Zhang C, Ma H, Wenjie Yan (2019) Lung Sound Recognition Algorithm Based on VGGish-biGRU. IEEE, pp 139438–139449
30. Yang X-K, Qu D, Zhang W-L, Zhang W-Q (2018) An adapted data selection for deep learning-based audio segmentation in the multi-genre broadcast channel. Digit. Sig. Process
31. Yin X, Liu C, Fang X (2021) Sentiment analysis based on BiGRU information enhancement. J Phys : Conf Ser

Kumari Nidhi Lal is working as an Assistant Professor at VNIT Nagpur. She received a B.tech degree in Information Technology in 2012 from Uttar Pradesh Technical University, Lucknow. Further, she received M.Tech Degree in 2015 in Information technology in the area of Wireless Communication and Computing from the Indian Institute of Information Technology in Allahabad, India. She completed her Ph.D. from the National Institute of Technology, Allahabad. She published many research papers in reputed scientific journals. Her current Research topics are Human-Computer Interaction, Multimedia analytics, and deep learning.