Time–Frequency Representation and Convolutional Neural Network-Based Emotion Recognition

Smith K. Khare and Varun Bajaj, *Senior Member, IEEE*

***Abstract*— Emotions composed of cognizant logical reactions toward various situations. Such mental responses stem from physiological, cognitive, and behavioral changes. Electroencephalogram (EEG) signals provide a noninvasive and nonradioactive solution for emotion identification. Accurate and automatic classification of emotions can boost the development of human–computer interface. This article proposes automatic extraction and classification of features through the use of different convolutional neural networks (CNNs). At first, the proposed method converts the filtered EEG signals into an image using a time–frequency representation. Smoothed pseudo-Wigner–Ville distribution is used to transform time-domain EEG signals into images. These images are fed to pretrained AlexNet, ResNet50, and VGG16 along with configurable CNN. The performance of four CNNs is evaluated by measuring the accuracy, precision, Mathew's correlation coefficient, F1-score, and false-positive rate. The results obtained by evaluating four CNNs show that configurable CNN requires very less learning parameters with better accuracy. Accuracy scores of 90.98%, 91.91%, 92.71%, and 93.01% obtained by AlexNet, ResNet50, VGG16, and configurable CNN show that the proposed method is best among other existing methods.**

***Index Terms*— Convolutional neural networks (CNNs), electroencephalogram (EEG), emotion recognition, smoothed pseudo-Wigner–Ville distribution (SPWVD).**

I. Introduction

EMOTION is a physiological state composed of expressive response, a physiological response, and subjective experience. In everyday life, emotions are important for engagement, interpretation, and decision-making. Human behavior, cognition, and communication are greatly influenced by emotions. The information regarding hobbies, interests, health, and so on may be interpreted by human emotions. Accurate recognition of emotions has emerged as a boon in the development of a human–computer interface [1]. Discrimination of human emotions can be accomplished by using facial expressions and speech [2], [3]. However, these models introduce the possibility of false classification because facial and speech expressions can be changed intentionally. By taking neurophysiological measurements, the problem can be overcome. Electroencephalogram (EEG) signals have gained copious attention due to its simplicity of acquisition and ease of use. EEG signals measure the electrical exercises of the brain and are hard to impact intentionally.

To date, researchers have proposed multiple methods for emotion classification based on EEG signals. Emotion classification from power spectral density (PSD), wavelet, and nonlinear dynamical features extraction using a support vector machine (SVM) has been proposed by Wang *et al.* [4]. The model proposed by them managed to achieve an accuracy of 83.55%. Nie *et al.* [5] used a series of bandpass filters, followed by the fast Fourier transform (FFT) to separate the chaotic rhythms. Log band energy has been classified by using SVM to attain an accuracy of 84.94%. Atkinson and Campos [6] used filtering-based rhythms separation to extract different feature sets. The dimensionality of features has been reduced by minimum redundancy, maximum relevance, and genetic algorithm (mRmR). The features have been classified by using SVM with an accuracy of 62.33%. Six time-domain features and five frequency-domain features extracted by FFT have been classified by using SVM, k-nearest neighbor (kNN), and multilayer perceptron (MLP) classifiers by Wang *et al.* [7]. The average accuracies of 59.84%, 63.07%, and 66.51% have been achieved by kNN, MLP, and SVM, respectively. Lin *et al.* [8] and Liu *et al.* [9] used a short-time Fourier transform (STFT) with nonoverlapping Hanning window for feature extraction. These features have been classified by using SVM and linear discriminant analysis (LDA). The model proposed by Lin *et al.* [8] and Liu *et al.* [9] achieved an accuracy of 80.86% and 65.09%, respectively. Ullah *et al.* [10] used multiple techniques based on STFT, discrete cosine transform, and spectrogram. These features have been classified by using SVM to identify human emotions with an accuracy of 73.5%. Lee and Lee [11] used STFT and convolutional neural network (CNN) with an accuracy of 89%.

Murugappan [12] used a wavelet transform (WT) with four different wavelets. Features extracted from the subbands have been classified by using kNN. An accuracy of 82.87% and 78.57% has been achieved with 62 and 24 channels. Murugappan *et al.* [13] and Mohammadi *et al.* [14] used discrete wavelet transform (DWT) to extract various features. These features have been classified by using kNN. The proposed method in [13] and [14] correctly predicted 83.26% and 86.75% of the emotions. Features based on statistical parameters, WT, and higher order crossing have been analyzed by using LDA, kNN, and SVM by Petrantonakis and Hadjileontiadis [15]. Their model achieved an accuracy of

Manuscript received February 4, 2020; revised April 30, 2020; accepted July 9, 2020. Date of publication July 31, 2020; date of current version July 7, 2021. *(Corresponding author: Varun Bajaj.)*

The authors are with the Electronics and Communication Discipline, Indian Institute of Information Technology Design and Manufacturing, Jabalpur 482005, India (e-mail: smith7khare@gmail.com; varunb@iiitdmj.ac.in).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.3008938

62.3% and 83.33% with LDA and SVM. Guo *et al.* [16] used the WT and fuzzy cognitive maps for emotion classification. Features extracted by using WT have been classified by hybrid SVM with an accuracy of 73.32%. Multiwavelet analysis with three multiwavelets, namely, Gernoimo-Hardin-Massopust, Chui Lian, and SA4, is presented by Bajaj and Pachori [17]. Features based on the Euclidian distance from phase space reconstruction have been classified by using a multiclass least square SVM (MC-LS-SVM). Zhuang *et al.* [18] used empirical mode decomposition (EMD) that extracts intrinsic mode functions (IMFs). Three features extracted from the IMF have been classified using SVM. The model proposed by Zhuang *et al.* [18] achieved an accuracy of 70.5%. Taran and Bajaj [19] used correlation-based filtering (CIF) method. The IMF and modes extracted by EMD and variational mode decomposition (VMD) have been filtered. Features from the modes have been classified by using MC-LS-SVM with an accuracy of 90.63%.

Bajaj *et al.* [20] used tunable Q wavelet transform (TQWT) to decompose the signal into low- and high-pass subbands. Features extracted from the subbands have been classified by the extreme learning machine (ELM). Their method provided 87.1% accurate separation of emotions. Bajaj *et al.* [21] and Gupta *et al.* [22] used flexible analytic wavelet transform (FAWT). Several features have been extracted and classified by using random forest and kNN. The method in [21] and [22] attains an accuracy of 86.1% and 87.5%, respectively. Separation of emotions using phase and angle reconstruction with Poincare feature extraction and SVM classification has been proposed in [23]. A hybrid deep belief network and hidden Markov model has been used to differentiate the emotions [24]. Differentiating the emotions by using self-organizing maps has been proposed in [25]. Common spatial patterns and PSD-based feature extraction methods have been used for emotion recognition using linear SVM [26]. Feature extraction based on asymmetric spatial patterns and Naive Bayes classifier has been used to identify the emotions [27]. The hybrid model based on the Hilbert–Huang spectrum, Zhao–Atlas–Marks distribution, and spectrogram methods has been used to classify emotions using SVM and kNN [28]. Quadratic time–frequency distribution and group sparse canonical correlation analysis have been used for the recognition of emotions [29], [30].

The methods based on filtering, FFT, and wavelet used an empirical selection of a type of filter, order, window, and wavelet. Choosing the length and type of window is an issue in STFT. The EMD-based method is purely experimental and lacks mathematical modeling. Due to the nonstationary nature of the EEG signals, an accurate selection of decomposition parameters of TQWT, FAWT, and VMD is difficult. The Zhao–Atlas–Marks distribution, Hilbert–Huang transform, and common spatial patterns are prone to noise. Moreover, the majority of the methods proposed in the literature have used manual feature extraction and classification methods. The traditional signal processing, feature extraction, and classification are time-consuming. These methods require huge qualitative and quantitative parameters analysis that greatly controls the performance of the system. Also, the methods in the literature are limited by its performance.

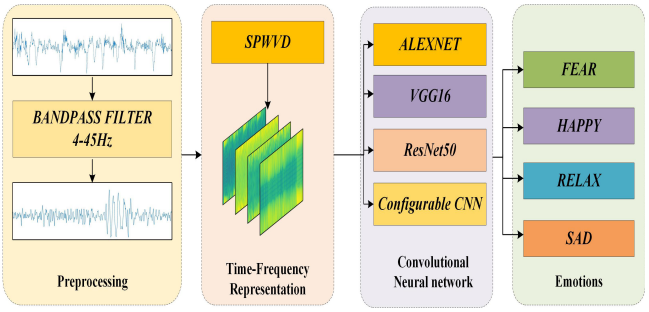

Fig. 1. Flowchart of the proposed framework.

The problem mentioned in the abovementioned literature creates an immediate need to develop automatic decomposition and classification of the signal. In this article, smoothed pseudo-Wigner–Ville distribution (SPWVD) and CNN-based emotions recognition is proposed. SPWVD is used for the transformation of a time-domain signal into time, frequency, and amplitude representation. The images of time–frequency representation (TFR) are given as an input to different CNNs. Three pretrained CNNs and a configurable CNN are used to classify the images. Several performance parameters are evaluated to get an insight into the proposed method. Finally, the superiority of the proposed method is tested by comparing it with the existing state of the art. The rest of this article is organized as follows. Section II describes the methodology of the proposed work. Results are discussed in Section III, and the discussion of the proposed method with existing methods is covered in Section IV. Finally, Section V presents the conclusion of the proposed method.

II. Methodology

A. Data-Set

Pictures, audio, video, and audio–video can be used to elicit various human emotions. The extraction of emotions from an audio–video method has outperformed other methods. The EEG recordings of 20 students with the mean age of 23 ± 0.5 years have been used. The data set is available online and the details of the experimental setup can be found in [19]–[21]. Audio–video clips of 10 s from Indian movies shown to the volunteers. The movie clips assumed to be self-explanatory, small, and intended to elicit single emotion. A 24-channel EEG recorder with transverse bipolar montage built according to the international 10–20 system has been used. The EEG signals have been recorded at a sampling frequency of 256 Hz. All the students considered for recordings do not have any physical or mental disorder. Four basic emotions, viz., fear, happy, relax, and sad, have been captured. Steps involved in the classification of emotions are shown in Fig. 1.

B. Preprocessing

The EEG signals are superimposed with noise sources that are not produced by neuronic actions called artifacts. EEG

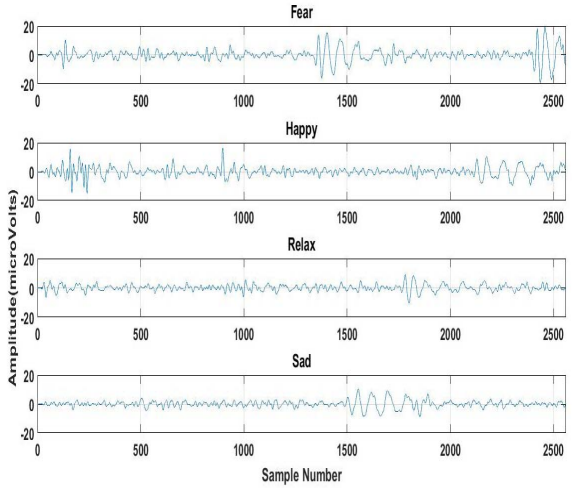

Fig. 2. Filtered EEG signals of four emotions.

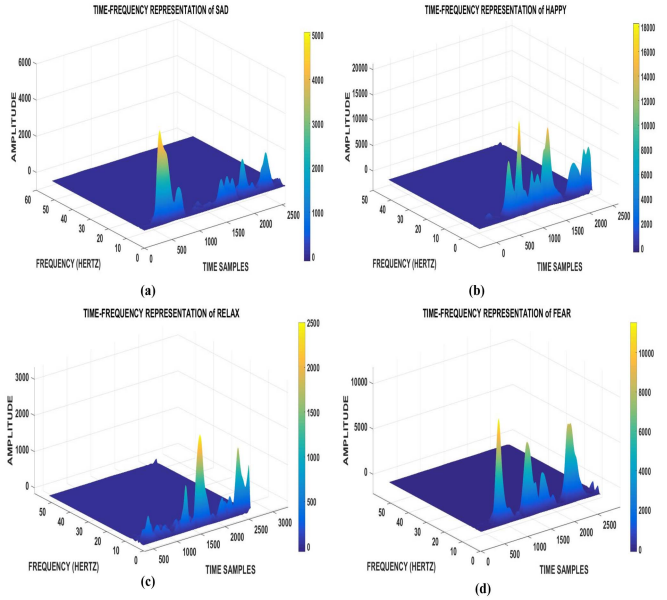

Fig. 3. TFR of EEG signals obtained by SPVWD. (a) Sad. (b) Happy. (c) Relax. (d) Fear.

signals are affected by the artifacts of retinal-corneo standing potentials. These potentials are measured between the back and the front portion of human eye called electrooculogram (EOG). The powerline signals of 50–60 Hz acts as noise to EEG signals. Frequencies useful for the identification of emotions are found below 40 Hz. The dominant frequency is selected, and artifacts are removed by preprocessing EEG signals. EEG signals are bandpass by using the tenth-order Butterworth filter. Passband and stopband frequencies are chosen to be 4 and 45 Hz, respectively [6], [31]. The frontal portion of human brain is significant for recording human reactions [32]. The EEG recordings of six frontal electrodes, namely, FP1, FP2, F3, F4, F7, and F8, have been used. As the electrodes are bipolar, the reading of FP2–F8, FP1–F3, FP2–F4, and FP1–F7 has been considered for signal processing and classification. The filtered EEG signals of fear, happy, relax, and sad are shown in Fig. 2. As seen from the figure, all the emotional states do not show any significant discriminative property of signals. Each signal contains a total of 2560 samples. There are four channels in total, as the bipolar montage has been used. Every channel of each class has 494 signals. A total of 1976 signals belong to each class.

C. Smoothed Pseudo-Wigner–Ville Distribution

CNNs require input as an image. Time-domain signals are converted into TFR to record the information in the spectral domain. A signal can be transformed into TFR by using STFT, Wigner–Ville distribution, SPWVD, continuous wavelet transform (CWT), and so on. TFR is a spatial representation of time, frequency, and amplitude simultaneously. TFR produced by STFT is known as a spectrogram. STFT requires the selection of window, its width, shape, and sampling frequency. This length must be maintained uniform throughout the signal. The spectrogram obtained by STFT gives poor resolution due to time–frequency localization. TFR obtained by CWT is called a scalogram. CWT needs the mother wavelet and its parameters to be chosen. The resolution of the scalogram depends on the choice of wavelet, while TFR obtained by the Wigner–Ville distribution produces cross term and attenuation for low frequency. To overcome these limitations, time-domain filtered EEG signals are transformed into a TFR by using SPWVD. SPWVD provides good time–frequency resolution that solves the problems of STFT and CWT. Limitations of the Wigner–Ville distribution is addressed by the introduction of cross-term reducing window in frequency domain. Hence, it is justified to choose SPWVD for signal transformation. SPWVD gives a direct representation of the time–frequency localization of signal energy. The length and type of cross-term reducing window in time and frequency domains can be chosen independently. Because of this, SPWVD provides good time–frequency cluster characteristics. The mathematical formulation of SPWVD can be represented by [33]

$$\varphi(t,f) = \int_{-\infty}^{+\infty} \gamma\left(t-t'\right)\phi\left(t',f\right)dt'$$
$$\phi\left(t',f\right) = \int_{-\infty}^{+\infty} h(\tau)z\left(t'+\frac{\tau}{2}\right)z^*\left(t'-\frac{\tau}{2}\right)e^{-j2\pi f t}d\tau \quad (1)$$

where $\gamma(t)$ and $h(t)$ are the cross-terms reducing windows in frequency and time domains. Time and frequency domain smoothing scales can be controlled easily. Length of the windows of $\gamma(t)$ and $h(t)$ can be selected independently. The TFR of filtered EEG signals obtained from SPWVD is shown in Fig. 3. As evident from the figure, all the states, namely, sad, happy, relax, and fear, show a discriminative representation of time–frequency–energy analysis. The energy amplitude of happy and fear state is very high (in the range of 10 000), energy amplitude of sad is medium (in the range of 5000), and energy amplitude of relax is low (in the range of 2500). The visual inspection of Figs. 2 and 3 shows that the transformed signals provide better insight information than the filtered time-domain EEG signals.

D. Convolutional Neural Networks

They are newly added subfield of machine learning domain. Inspired by artificial neural networks, CNN is comprised of self-optimized neurons. CNN is also known as a deep learning network that automatically classifies the signals. Inspired by mice's visual system, CNN is designed to work with images. CNN takes the spacial and configural structure of the input into account [34].

Recently, CNN is one of the most widely used techniques for image classification, object detection, face recognition, and so on. CNN is composed of multilayers interconnected neurons trained rigorously for feature extraction and classification. CNN replaces the time-consuming traditional feature extraction and classification algorithms. CNN learns automatically to extract the features and to classify them. Because of its transfer and automated learning characteristics, CNN finds a humungous application in computer vision. The system is trained on one task and reused for other tasks. CNN is comprised of an input layer, multiple hidden layers, and an output layer. A hidden layer of CNN is composed of a convolutional layer (CL), a pooling layer (PL), and a fully connected (FC) layer. Extraction of high-level features is carried out by CL and PL. The classification task is governed by FC layers. The function of each layer is explained as follows.

1) *CL:* It is the key that decides the operation of CNN. The performance of CNN depends on the use of learnable filters. Spatial dimensionality of the kernels is usually small but spread along with the entire depth of the image. The 2-D convolution of the signal with two dimensions can be written as

$$(M * N)(m, n) = \sum_{i,j} M(i, j)N(m + i, n + j). \quad (2)$$

The filters are usually moved by the number of pixels called stride (q). Sometimes, maintaining the size of the image zero padding may also be applied with size z. For an image input with dimension, $W_m \times H_m \times K_m$, where W_m is the width, H_m is the height, and K_m are the number of channels. With K_0 filters each of size $r \times r$, the output volume $W_0 \times H_0 \times K_0$ can be written as

$$W_0 = \frac{W_m - r + 2z}{q} + 1$$
$$H_0 = \frac{H_m - r + 2z}{q} + 1. \quad (3)$$

The operation of convolution is combined with an activation function. The activation function enhances the nonlinearity in the network. The most common activation function used is the rectified linear unit (ReLu).

2) *PL:* CL is succeeded by the PL also known as subsampling layer. The main objective of PL is to produce downsampled feature maps. It reduces the parameters and dimensions by keeping useful information. The PL also helps to regulate overfitting. It operates on each activation map by using max or mean functions. For J input maps, the output maps are generally smaller as given by

$$x_k^l = f\left(\alpha_k^l \text{down}\left(x_k^{l-1}\right) + \beta_k^l\right) \quad (4)$$

where α_k^l and β_k^l are the multiplicative and additive bias terms and down$(\cdot)$ is the pooling function. The output of PL is given as an input to an FC layer.

3) *FC Layer:* Pooling layer is succeeded by an FC layer. It is a feedforward neural network. FC converts a 2-D feature map into a 1-D feature map. The softmax layer converts the score into probabilities, and at last, based on some algorithm, the classification layer assigns a class to an object.

Using the abovementioned layers, one can build their own CNN. The number of convolutional, pooling, and FC layers can be added or dropped until the desired performance of the network is obtained. With recent advances in CNN, many pretrained deep CNNs have been used for various machine learning problems. AlexNet, ResNet50, VGG16, VGG19, GoogleNet, and so on are some of the well-known pretrained transfer learning networks. These networks transfer the previously learned knowledge of one domain to another for feature extraction and classification. New images are used for training with fewer numbers as used in the previously trained data set. In this work, three benchmark CNNs, namely, AlexNet, ResNet50, and VGG16, are used for emotion recognition. The details of these networks can be found in [35]. There is no standard CNN method available for the analysis and classification of EEG signals. The choice of CNN depends on the performance given by it. Many existing CNNs have a large number of layers. Complex architecture increases the number of learnable parameters significantly. Moreover, the time required for training, testing, and validation is higher for complex networks. The performance of CNN highly depends on the hyperparameters. By varying the filter size, stride, dropout, and so on, classification accuracy can be varied. Better accuracy may be achieved with fewer parameters and lesser complexity [36], [37]. Motivated by this fact, a configurable CNN with a fewer number of a CL, pooling, and size of the FC layer is designed. The configurable network consists of four CL, two PL, a dropout, and two FC layers. The architecture of the proposed network can be modified according to the application. The number of CLs and PLs can be added or deleted as per user choice. Also, the number of learnable parameters required for the proposed architecture is less. The architecture of the configurable CNN is shown in Fig. 4.

III. Results

The traditional classification problems involve signal decomposition, feature extraction, feature selection, and classification. The performance of the conventional systems depends highly on parameters selected for decomposition and classification. The signal analysis and classification using conventional methods are time-consuming and laborious. With this motivation, an automatic and reliable classification of emotions is proposed in this work. EEG signals of four emotions elicited from the audio–video clips are used. The artifacts and noise are removed by filtering the EEG signals using a bandpass filter. The filtered 1-D EEG signals are converted into TFR using SPWVD. The images obtained are given as an input to three benchmark pretrained networks and configurable CNN with four CLs and two FC layers.

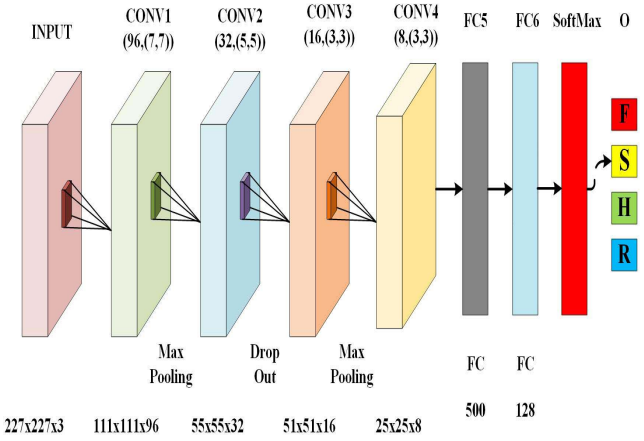

Fig. 4. Network architecture of configurable CNN.

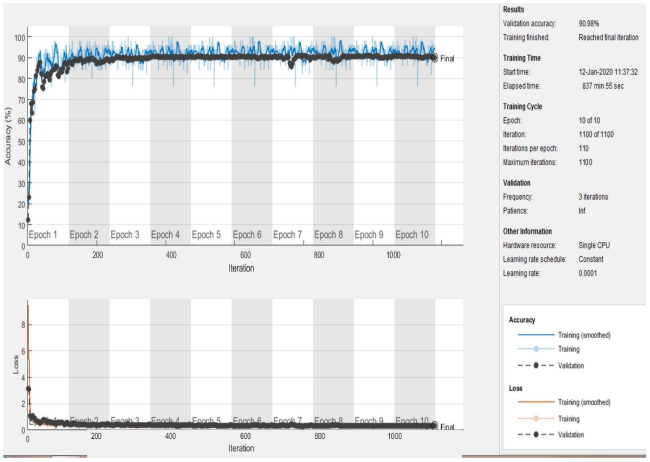

Fig. 5. Classification accuracy and loss of Alexnet.

Raw EEG signal is given to the tenth-order Butterworth filter. The passband frequency of 4–45 Hz is selected with a sampling frequency 256 Hz. The filtered signal obtained from the BPF is given as an input to TFR. In this methodology, SPWVD is used to convert the 1-D signal into a 2-D signal. The Kaiser window is used for reducing the cross terms in time and frequency domains. Too small window size may result in poor resolution and too large windows might increase the size of the image drastically. Hence, medium-size window with length 31 is chosen empirically. The window size is kept $2^n - 1$ for fast computation, where n is the number of bits. The TFR obtained from SPWVD is fed to AlexNet, ResNet50, VGG16, and proposed network. Multiple networks are employed as the performance of one learning algorithm can be overthrown by others due to lack of priori as stated in "No free lunch theorem" [38].

The common experimental platform is maintained for training and testing of all the networks; 70% data set is used for training the network and the rest is used for testing. Weight and bias learning rate is fixed to 20. Adam optimizer is used for scaling the learning rate for each weight of the neural network. The batch size and number of epochs are fixed to 50 and 10, respectively. The learning rate is fixed to 0.0001 and the validation frequency is set to 3. A total of 1100 iterations are carried out with 110 iterations per epoch. AlexNet is an eight-layer network with five CLs and three FC layers. AlexNet takes input images size with dimensions 227×227. Convolution and max pooling with local response normalization are performed in the first CL, 96 filters each with size 11×11 and max pooling of size 3×3 with a stride of 2. Second, CL composed of 256 receptive filters each of 5×5. The third and fourth layers comprise 384 feature maps each of 3×3 filters. The fifth layer has 296 filters each of size 3×3. Sixth and seventh are two FC layers succeeded by the dropout and softmax layer. The overall validation accuracy obtained by using AlexNet is 90.98%. Fig. 5 shows the training and validation values of accuracy and loss per iteration. The time required to reach the final iteration is 837 min and 55 s.

Table I represents the confusion matrix obtained from AlexNet. The fear state is 96.91% accurately classified with very less misclassification of 2.38%, 0.35%, and 0.66% in happy, relax, and sad states, respectively. Classification accuracy of happy, relax, and sad emotional state is 88.77%, 83.45%, and 95.09%, respectively.

TABLE I
CONFUSION MATRIX OF ALEXNET

Class	Fear	Happy	Relax	Sad
Fear	96.61	2.38	0.35	0.66
Happy	2.18	88.77	4.55	4.50
Relax	4.55	3.14	83.45	8.86
Sad	1.37	1.92	1.62	95.09

TABLE II
CONFUSION MATRIX OF RESNET50

Class	Fear	Happy	Relax	Sad
Fear	95.70	2.07	0.76	1.47
Happy	1.77	87.04	0.86	10.32
Relax	2.48	1.27	90.99	5.26
Sad	1.11	2.88	2.07	93.93

The ResNet50 network consists of 50 CLs and single FCs. The filters size of ResNet50 is 1×1, 3×3, and 7×7. The input size of an image taken by this network is of 224×224. The resized images of 224×224 obtained by SPWVD are given as input. Accuracy and loss per iteration for ResNet50 are shown in Fig. 6. The obtained accuracy of ResNet50 is 91.91% with the total time required for testing and validation is 3325 min and 50 s. The confusion matrix obtained using ResNet50 is shown in Table II. As evident from the table, the classification accuracy of 95.70%, 87.04%, 90.99%, and 93.93% is obtained for fear, happy, relax, and sad states, respectively.

Another well-known and popular pretrained network, namely VGG16, is used to test the system performance. VGG16 takes input of image size 224×224. It consists of 16 CLs and three FC layers with a filter size of 3×3. The classification accuracy of VGG16 is obtained as 92.71% with total time required for training and testing is 2320 min

Fig. 6. Classification accuracy and loss of ResNet50.

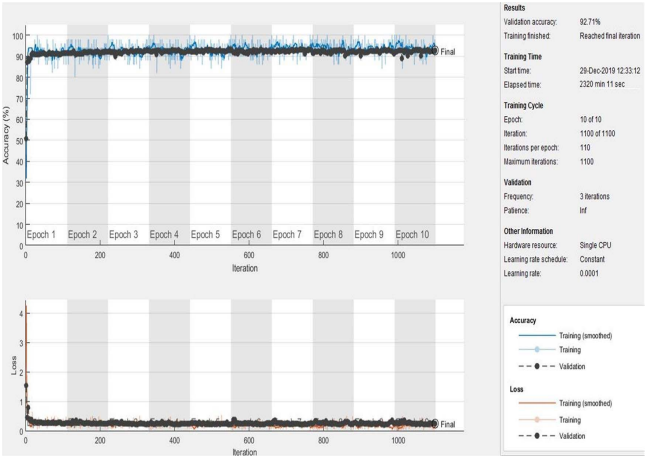

Fig. 7. Classification accuracy and loss of VGG16.

TABLE III
CONFUSION MATRIX OF VGG16

Class	Fear	Happy	Relax	Sad
Fear	97.06	1.62	0.56	0.76
Happy	2.73	87.25	5.72	4.30
Relax	1.21	2.68	93.17	2.94
Sad	2.58	1.97	2.07	93.37

and 11 s. Validation accuracy and loss are shown in Fig. 7. The confusion matrix of VGG16 is shown in Table III. Classification accuracy of 97.06% is obtained for fear with 1.62%, 0.56%, and 0.76% misclassification in happy, relax and sad states, respectively. Happy and relax states have a classification accuracy of 87.25% and 93.17%. Misclassification obtained for a fear state is 2.73%, 2.58%, and 1.21%. Misclassification for relax is 5.72% and 2.07%, misclassification for happy state is 2.68% and 1.97%, and misclassification for sad state is 4.30% and 2.94%. The classification accuracy of the sad state is 93.37%.

Finally, configurable CNN is used for testing the performance. It consists of four CLs, two PLs, and two FC layers

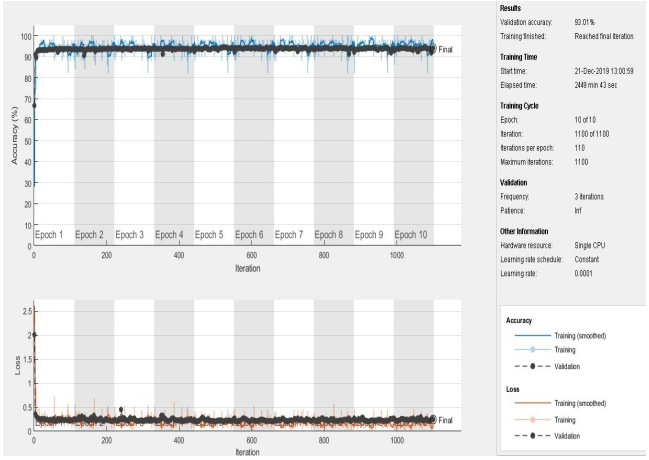

Fig. 8. Classification accuracy and loss of configurable CNN.

TABLE IV
CONFUSION MATRIX OF CONFIGURABLE CNN

Class	Fear	Happy	Relax	Sad
Fear	96.71	0.61	1.27	1.42
Happy	0.76	86.08	4.76	8.40
Relax	0.40	0.61	93.83	5.16
Sad	0.30	3.44	0.81	95.45

having a dropout of 50% with a filter size of 3×3, 5×5, and 7×7. This network takes input images with a size of 227×227 and Adam optimization for learning the weights. The training and validation accuracy per iteration is shown in Fig. 8. An accuracy of 93.01% is achieved having a total time of 2449 min and 43 s. The confusion matrix of configurable CNN is shown in Table IV; 96.71%, 86.08%, 93.83%, and 95.45% accurate classification are obtained for fear, happy, relax, and sad states, respectively. Misclassification for fear is 0.76%, 0.40%, and 0.30%. Misclassification obtained for happy state is 0.61%, 0.61%, and 3.44%.

Effectiveness of the proposed method is tested by evaluating different performance parameters. Table V shows five performance parameters, namely accuracy, precision, Mathew's correlation coefficient (MCC), F1-score, and false-positive rate (FPR) obtained for different CNNs. For AlexNet, accuracy, precision, MCC, F1, and FPR of 90.98%, 91.11%, 87.85%, 0.9092, and 3.2% are obtained, respectively. ResNet50 and VGG16 provide the accuracy of 91.91% and 92.71%, the precision of 92.19% and 92.72%, the MCC of 89.21% and 90.17%, the F1-score of 0.9195 and 0.9268, and FPR of 2.82% and 2.55%, respectively. Configurable CNNs mark as the accuracy of 93.01%, the precision of 93.26%, MCC of 90.70%, F1-score of 0.9302, and FPR of 2.42%.

Deeper is better? To answer this question, a comparison of different networks is made in Table VI. The number of CL in a configurable CNN is 4, which is fewer compared with benchmark networks. An FC layer with a fever number of neurons is used that reduces the computational complexity significantly. The CL is designed with a filter size of 3, 5, and 7 compared

TABLE V
PERFORMANCE PARAMETERS OF DIFFERENT NETWORKS

Parameters	AlexNet	ResNet50	VGG16	Configurable
Accuracy	90.98	91.91	92.71	93.01
Precision	91.11	92.19	92.72	93.26
MCC	87.85	89.21	90.17	90.7
F1-Measure	0.9092	0.9195	0.9268	0.9302
FPR	3.2	2.82	2.55	2.42

TABLE VI
DETAILS OF PARAMETERS FOR DIFFERENT CNN

Parameters	AlexNet	ResNet50	VGG16	Configurable
CL	5	50	16	4
FCL	3	1	3	2
Filter Size	3,5,11	1,3,7	3	3,5,7
Approx. Time (min)	838	3326	2320	2450
No. Parameters	61M	25.5M	138M	0.75M
Stride	1,4	1,2	1	1,2
Accuracy	90.98	91.91	92.71	93.01

with 3, 5, and 11 for AlexNet, 1, 3, and 7 for ResNet50, and 3 in the case of VGG16. Learnable parameters required in the case of AlexNet are 61 million (M), whereas for ResNet50 and VGG16, the parameters are 25.5M and 138M, respectively. With 737 452 number of learnable parameters, complexity is reduced significantly in configurable CNN. The total time required for training AlexNet, ResNet50, and VGG16 is approximately 838, 3326, and 2320 min, respectively. The time required to train the configurable CNN is 2450 min that is higher than AlexNet and VGG16 and lower than ResNet50. The accuracy of the configurable CNN is higher than all the benchmark networks used.

IV. DISCUSSION

The proposed system is tested for efficacy by comparing it with other existing methods, as shown in Table VII. Their methods have been applied to the data set used in the proposed method. The same decomposition and classification techniques have been employed to test the system efficacy. Bajaj *et al.* [21] used FAWT to decompose the signals into subbands. Six time-domain features elicited from eight subbands have been used as an input to different kernels of KNN. The accuracy achieved with weighted KNN reported highest with a value of 86.1%. The method proposed in [13] used DWT to extract several features from the subbands. The features have been classified by using kNN. Their method managed to provide an 82.32% accurate system for emotion identification. Wang *et al.* [7] extracted multiple features in the time and frequency domains. Six time-domain statistical features and five frequency-domain features have been selected with mRmR. These features have been classified by using kNN, MLP, and SVM. Their method achieved a mean accuracy

TABLE VII
PERFORMANCE COMPARISON WITH RESPECT TO CLASSIFICATION ACCURACY WITH EXISTING STATE OF THE ART

Authors	Used Method	Emotions	Accuracy
Bajaj et. al [21]	FAWT & kNN	4	86.1
Murugappan et. al [13]	DWT & kNN	4	82.32
Wang et. al. [7]	FFT & kNN	4	62.58
	FFT & MLP		64.25
	FFT & SVM		71.26
Lin et. al. [8]	STFT & SVM	4	76.48
Bajaj et. al [20]	TQWT & ELM	4	87.1
Taran et. al [19]	CIF & MC-LS-SVM (MH)	4	86
	CIF & MC-LS-SVM (Polynomial)		88.66
	CIF & MC-LS-SVM (RBF)		88.78
	CIF & MC-LS-SVM (Morlet)		90.63
Proposed	**SPWVD & AlexNet**	**4**	**90.98**
	SPWVD & ResNet50		**91.91**
	SPWVD & VGG16		**92.71**
	SPWVD & configurable CNN		**93.01**

of 62.58%, 64.25%, and 71.26% with kNN, MLP, and SVM on this data set, respectively. The features extracted using STFT with a Hanning window in the method proposed by Lin *et al.* [8] have been classified by using SVM. Their model achieved an accuracy of 76.48% with the same data set. Another method proposed by Bajaj *et al.* [20] employed TQWT. Their method uses eight subbands to evaluate time-domain features. Ten time-domain features have been extracted from the subbands of four frontal channels. The extracted features are given as an input to ELM that gives an accuracy of 87.1%. Taran *et al.* [19] employed two-stage CIF. EMD and VMD have been used to filter the dominant subbands of the EEG signals.

This method used a single-channel feature evaluation. The extracted features have been classified by using four kernels of MC-LS-SVM. Mexican hat, polynomial, radial basis function, and Morlet kernel produce an accuracy of 86%, 88.66%, 88.78%, and 90.63%, respectively. The method proposed in this article uses four different CNN architectures. TFR obtained by using SPWVD is fed to AlexNet, ResNet50, VGG16, and configurable CNN. An accuracy of 90.98% is obtained by using AlexNet. Classification accuracy of 91.91% is accomplished when employing ResNet50. VGG16 architecture classifies 92.71% images accurately. The configurable-designed CNN with four CLs and two FC layers marks an accuracy of 93.01%. As evident from the table, pretrained AlexNet, VGG16, and ResNet50 architectures, and the configurable CNN outperform other states of the art. The advantages and limitations of the proposed method are listed as follows.

Advantages:

1) Reliability and simplicity.
2) Method is tunable as per the application.
3) Robust in terms of its scope with other transformation techniques and data sets.

Limitations:

1) Use of empirical parameters for signal processing and classification.
2) Testing and validation are performed on a single data set.

V. Conclusion

Various CNNs are examined in this article to categorize four emotions using EEG signals. Compared with the traditional approach, CNN has added edge in terms of automatic extraction and classification of features. The approach presented in this article uses the filtering technique and the SPWVD to transform time-domain EEG input signals into images. TFR obtained by SPWVD of four basic emotions, namely, fear, happy, relax, and sad, are given to four CNNs. Performance of three pretrained networks is AlexNet, ResNet50, and VGG16, and a configured CNN with four CLs and two FC layers are compared. The results obtained through these networks show that AlexNet offers quicker training and testing. VGG16 offers second-fastest training, whereas ResNet50 is slowest. Configurable CNN provides maximum precision over pretrained networks with significantly fewer learnable parameters. The comparative results demonstrate the superiority of the new approach over the existing methods. This methodology can be used in the development of EEG-based human–computer interface. An optimal selection of windows and their size can be used in the future to convert EEG signals into an image. Hyperparameter optimization can be explored to improve the performance of the system. In the near future, the proposed method will be tested on other data sets and different biopotential signals.

References

[1] Z. Yin, Y. Wang, L. Liu, W. Zhang, and J. Zhang, "Cross-subject EEG feature selection for emotion recognition using transfer recursive feature elimination," *Frontiers Neurorobot.*, vol. 11, p. 19, Apr. 2017.

[2] M. Song, C. Chen, J. Bu, and M. You, "Speech emotion recognition and intensity estimation," in *Computational Science and Its Applications—ICCSA*. Berlin, Germany: Springer, 2004, pp. 406–413.

[3] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 757–763, Jul. 1997.

[4] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, Apr. 2014.

[5] D. Nie, X.-W. Wang, L.-C. Shi, and B.-L. Lu, "EEG-based emotion recognition during watching movies," in *Proc. 5th Int. IEEE/EMBS Conf. Neural Eng.*, Apr. 2011, pp. 667–670.

[6] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Syst. Appl.*, vol. 47, pp. 35–41, Apr. 2016.

[7] X.-W. Wang, D. Nie, and B.-L. Lu, "EEG-based emotion recognition using frequency domain features and support vector machines," in *Neural Information Processing*, B.-L. Lu, L. Zhang, and J. Kwok, Eds. Berlin, Germany: Springer, 2011, pp. 734–743.

[8] Y.-P. Lin *et al.*, "EEG-based emotion recognition in music listening," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 7, pp. 1798–1806, Jul. 2010.

[9] Y.-J. Liu, M. Yu, G. Zhao, J. Song, Y. Ge, and Y. Shi, "Real-time movie-induced discrete emotion recognition from EEG signals," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 550–562, Oct. 2018.

[10] H. Ullah, M. Uzair, A. Mahmood, M. Ullah, S. D. Khan, and F. A. Cheikh, "Internal emotion classification using EEG signal with sparse discriminative ensemble," *IEEE Access*, vol. 7, pp. 40144–40153, 2019.

[11] H.-J. Lee and S.-G. Lee, "Arousal-valence recognition using CNN with STFT feature-combined image," *Electron. Lett.*, vol. 54, no. 3, pp. 134–136, Feb. 2018.

[12] M. Murugappan, "Human emotion classification using wavelet transform and KNN," in *Proc. Int. Conf. Pattern Anal. Intell. Robot.*, vol. 1, Jun. 2011, pp. 148–153.

[13] M. Murugappan, N. Ramachandran, and Y. Sazali, "Classification of human emotion from EEG using discrete wavelet transform," *J. Biomed. Sci. Eng.*, vol. 3, no. 4, pp. 390–396, 2010.

[14] Z. Mohammadi, J. Frounchi, and M. Amiri, "Wavelet-based emotion recognition system using EEG signal," *Neural Comput. Appl.*, vol. 28, no. 8, pp. 1985–1990, Aug. 2017.

[15] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from EEG using higher order crossings," *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 186–197, Mar. 2010.

[16] K. Guo *et al.*, "A hybrid fuzzy cognitive map/support vector machine approach for EEG-based emotion classification using compressed sensing," *Int. J. Fuzzy Syst.*, vol. 21, no. 1, pp. 263–273, Feb. 2019.

[17] V. Bajaj and R. B. Pachori, "Human emotion classification from EEG signals using multiwavelet transform," in *Proc. Int. Conf. Med. Biometrics*, May 2014, pp. 125–130.

[18] N. Zhuang, Y. Zeng, L. Tong, C. Zhang, H. Zhang, and B. Yan, "Emotion recognition from EEG signals using multidimensional information in EMD domain," *BioMed Res. Int.*, vol. 2017, Aug. 2017, Art. no. 8317357.

[19] S. Taran and V. Bajaj, "Emotion recognition from single-channel EEG signals using a two-stage correlation and instantaneous frequency-based filtering method," *Comput. Methods Programs Biomed.*, vol. 173, pp. 157–165, May 2019.

[20] A. H. Krishna, A. B. Sri, K. Y. V. S. Priyanka, S. Taran, and V. Bajaj, "Emotion classification using EEG signals based on tunable-Q wavelet transform," *IET Sci., Meas. Technol.*, vol. 13, no. 3, pp. 375–380, May 2019.

[21] V. Bajaj, S. Taran, and A. Sengur, "Emotion classification using flexible analytic wavelet transform for electroencephalogram signals," *Health Inf. Sci. Syst.*, vol. 6, no. 1, p. 12, Sep. 2018.

[22] V. Gupta, M. D. Chopda, and R. B. Pachori, "Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals," *IEEE Sensors J.*, vol. 19, no. 6, pp. 2266–2274, Mar. 2019.

[23] M. Zangeneh Soroush, K. Maghooli, S. K. Setarehdan, and A. M. Nasrabadi, "A novel EEG-based approach to classify emotions through phase space dynamics," *Signal, Image Video Process.*, vol. 13, no. 6, pp. 1149–1156, Sep. 2019.

[24] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu, "EEG-based emotion classification using deep belief networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2014, pp. 1–6.

[25] R. Khosrowabadi, H. C. Quek, A. Wahab, and K. K. Ang, "EEG-based emotion recognition using self-organizing map for boundary detection," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 4242–4245.

[26] S. Koelstra *et al.*, "Single trial classification of EEG and peripheral physiological signals for recognition of emotions induced by music videos," in *Brain Informatics*, Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong, and J. Huang, Eds. Berlin, Germany: Springer, 2010, pp. 89–100.

[27] D. Huang, C. Guan, K. Keng Ang, H. Zhang, and Y. Pan, "Asymmetric spatial pattern for EEG-based emotion detection," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–7.

[28] S. K. Hadjidimitriou and L. J. Hadjileontiadis, "Toward an EEG-based recognition of music liking using time-frequency analysis," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3498–3510, Dec. 2012.

[29] R. Alazrai, R. Homoud, H. Alwanni, and M. Daoud, "EEG-based emotion recognition using quadratic time-frequency distribution," *Sensors*, vol. 18, no. 8, p. 2739, Aug. 2018.

[30] W. Zheng, "Multichannel EEG-based emotion recognition via group sparse canonical correlation analysis," *IEEE Trans. Cognit. Develop. Syst.*, vol. 9, no. 3, pp. 281–290, Sep. 2017.

[31] M. S. Özerdem and H. Polat, "Emotion recognition based on EEG features in movie clips with channel selection," *Brain Informat.*, vol. 4, no. 4, pp. 241–252, Dec. 2017.

[32] Y.-Y. Lee and S. Hsieh, "Classifying different emotional states by means of EEG-based functional connectivity patterns," *PLoS ONE*, vol. 9, no. 4, pp. 1–13, Apr. 2014.

[33] E. P. de Souza Neto, M.-A. Custaud, J. Frutoso, L. Somody, C. Gharib, and J.-O. Fortrat, "Smoothed pseudo Wigner–Ville distribution as an alternative to Fourier transform in rats," *Autonomic Neurosci.*, vol. 87, nos. 2–3, pp. 258–267, Mar. 2001.

[34] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[35] A M. Zahangir *et al.*, "The history began from AlexNet: A comprehensive survey on deep learning approaches," Mar. 2018, *arXiv:1803.01164*. [Online]. Available: https://arxiv.org/abs/1803.01164

[36] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size," Feb. 2016, *arXiv:1602.07360*. https://arxiv.org/abs/1602.07360
[37] S. Han, H. Mao, and W. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," Oct. 2015, *arXiv:1510.00149*. [Online]. Available: https://arxiv.org/abs/1510.00149
[38] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1341–1390, Oct. 1996.

Smith K. Khare received the B.E. degree from the Shri Ramdeobaba College of Engineering and Management, Nagpur, India, in 2012, and the M.Tech. degree from Mumbai University, Mumbai, India, in 2015.

He was working as an Assistant Professor with the Yeshwantrao Chavan College of Engineering, Nagpur, India, the G. H. Raisoni College of Engineering, Nagpur, and the Shri Ramdeobaba College of Engineering and Management, Nagpur. He is currently a Research Scholar with the PDPM-Indian Institute of Information Technology, Design and Manufacturing, Jabalpur, India. He has authored or coauthored ten publications in various high impact factor, peer-reviewed journals, such as IEEE TRANSACTIONS and Elsevier. He has published two papers in international conference. His research interests include biomedical signal processing, pattern recognition, machine learning, and deep neural networks.

Mr. Khare is serving as a reviewer for IEEE, Elsevier, and several other reputed journals.

Varun Bajaj (Senior Member, IEEE) received the B.E. degree in electronics and communication engineering from Rajiv Gandhi Technological University, Bhopal, India, in 2006, the M.Tech. degree (Hons.) in microelectronics and VLSI design from the Shri Govindram Seksaria Institute of Technology and Science, Indore, India, in 2009, and the Ph.D. degree from the Discipline of Electrical Engineering, IIT Indore, Indore, India, in 2014.

He worked as a Visiting Faculty Member with the Indian Institute of Information Technology, Design and Manufacturing (IIITDM), Jabalpur, India, from September 2013 to March 2014, where he has been working as a Faculty Member with the Discipline of Electronics and Communication Engineering since 2014. He was an Assistant Professor with the Department of Electronics and Instrumentation, Shri Vaishnav Institute of Technology and Science, Indore, from 2009 to 2010. He has edited *Modelling and Analysis of Active Biopotential Signals in Healthcare* (IOP Books, Volumes 1 and 2). He has authored more than 100 research papers in various reputed international journals/conferences, such as IEEE TRANSACTIONS, Elsevier, Springer, and IOP. The citation impact of his publications is around 1800 citations, H-index of 19, and i10 index of 40 (Google Scholar July 2020). He has guided three (03) Ph.D. scholars and five M.Tech. scholars. His research interests include biomedical signal processing, image processing, time–frequency analysis, and computer-aided medical diagnosis.

Dr. Bajaj was a recipient of various reputed national and international awards. He has served as a Subject Editor for *IET Electronics Letters* from November 2018 to June 2020, where he is also serving as a Subject Editor-in-Chief. He is also contributing as an active technical reviewer for leading International journals of IEEE, IET, Elsevier, and so on.