IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 25, 2021, accepted September 7, 2021, date of publication September 10, 2021,
date of current version September 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3111659

Speech Emotion Recognition Using Clustering Based GA-Optimized Feature Set

SOFIA KANWAL[1,2] AND SOHAIL ASGHAR[2], (Member, IEEE)
[1]Department of CS and IT, University of Poonch Rawalakot, Azad Kashmir 12350, Pakistan
[2]Department of Computer Science, COMSATS University Islamabad, Islamabad Campus, Islamabad 45550, Pakistan
Corresponding author: Sofia Kanwal (sofiakanwal@upr.edu.pk)

ABSTRACT Speech Emotion Recognition (SER) is a hot topic in academia and industry. Feature engineering plays a pivotal role in building an efficient SER. Although researchers have done a tremendous amount of work in this field, there are still the issues of speech feature choice and the correct application of feature engineering that remains to be solved in the domain of SER. In this research, a feature optimization approach that uses a clustering-based genetic algorithm is proposed. Instead of randomly selecting the new generation, clustering is applied at the fitness evaluation level to detect outliers for exclusion to be part of the next generation. The approach is compared with the standard Genetic Algorithm in the context of audio emotion recognition using Berlin Emotional Speech Database (EMO-DB), Ryerson Audio-Visual Database of Speech and Song (RAVDESS) and, Surrey Audio-Visual Expressed Emotion Dataset (SAVEE). Results signify that the proposed technique effectively improved the emotion classification in speech. The recognition rate of 89.6% for general speakers (both male and female), 86.2% for male speakers, and 88.3% for female speakers on EMO-DB, 82.5% for general speakers, 75.4% for male speakers, and 91.1% for female speaker on RAVDESS, and 77.7% for general speakers on SAVEE is obtained in speaker-dependent experiments. For speaker-independent experiments, we achieved the recognition rate of 77.5% on EMO-DB, 76.2% on RAVDESS and, 69.8 % on SAVEE. All the experiments were performed on MATLAB and the Support Vector Machine (SVM) was used for classification. Results confirm that the proposed method is capable of discriminating emotions effectively and performed better than the other approaches used for comparison in terms of performance measures.

INDEX TERMS Clustering, feature engineering, feature optimization, genetic algorithm, OpenSMILE tool kit, speech emotions, support vector machine.

I. INTRODUCTION

Speech is one of the most widely used and direct ways of conveying emotions and perceiving the feelings of others [1]. This is why the success of many applications such as auto-replies, chat-bot, speaking humanoid robots and many other scenarios involving human-machine interaction are dependent on speech emotion recognition. The emotion analysis of speech signals has been studied for the last two decades [2]. Currently, there are many models that use machine learning and deep learning for audio emotion recognition [3]–[6], [7]. Features are extracted and classification is performed. The effective classification performance relies on the quality and quantity of features being used. In this respect, feature engineering is an important stage in classification task.

The associate editor coordinating the review of this manuscript and approving it for publication was M. Shamim Hossain.

Primarily the classification performance in terms of accuracy and time to calculate the accuracy is heavily dependent on how the features are being engineered i.e. the type of features being selected, the quantity of selected features, and the type of feature optimization employed [8]. All this is essential to recognize emotions robustly across datasets of different languages, different sizes, and different emotion ranges. Along with the global optimum solutions proposed in literature [9], [10] [11], there exists optimization solutions with the specific focus on feature optimization in SER [12], [13]. Even though, these methods improved the accuracy of the SER systems, however, there is still a great margin to achieve.

The optimization algorithms are of two types: Deterministic and Stochastic. Most of the classical optimization strategies used are deterministic. However, for problems that have a variety of sub-problems, stochastic algorithms are preferred

to use [14]. The objective of optimization algorithms is to find the best possible solutions within a feasible time limit. In one of the studies [15], Yogesh *et al.* presented a new particle swarm-based optimization for feature selection in SER. To obtain the optimum values of features for speech emotions, Gharwan has used an optimization algorithm [11]. In another study [16], deep convolution neural networks are used to extract spectrogram features. These algorithms do not guarantee the best solution, however, most of the times nearly optimal solution is found.

Genetic Algorithms (GA) are optimization techniques [17]. In several studies, GA is giving promising results when used for feature engineering [15], [18], [19]. Thus, in this research, we are employing the GA for feature optimization in speech emotion recognition.

Along with feature optimization and selection, classification phase is equally important to build a robust SER system. Recent studies are more focused on deep learning approaches for feature engineering and classification [7], [16], [20], [21]. However, in most of the studies involving SER, classifiers such as SVM [22], decision trees [23], and k-nearest neighbor (kNN) [24] are used as single, multiple, hybrid, or ensemble forms [25]. In this study, we are using SVM which is one of the standard algorithms and most frequently used classifier [26].

The overall objective of this research is to improve emotion recognition performance in terms of accuracy. For this, we have proposed clustering-based GA for feature optimization. Clustering is the way of organizing data into groups based on similar features. Clustering has many applications like pattern recognition [27], image analysis [28], and data mining [29]. We are using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for outlier detection at the feature optimization level.

We name the newly proposed optimization algorithm as Density-Based Spatial Clustering of Application with Noise Genetic Algorithm (DGA). The performance of (DGA) is evaluated on optimization of speech emotion features using three datasets. The numerical results reveal its superiority over the standard Genetic Algorithm. It is composed of five stages: generation of initial population, clustering, detecting outliers, selection of parents from non-outliers for cross-over and mutation.

A. CONTRIBUTION OF THIS RESEARCH

This section defines the contribution of our research concretely. We have proposed a feature optimization algorithm that uses clustering-based GA. In GA, we have used clustering mechanism for fitness evaluation. Feature optimization is performed using the proposed method. Our method outperformed the standard genetic algorithm which uses a random selection for fitness evaluation. The accuracy gain of 1.9%, 14.87% and, 1.36% for general, male, and female tests respectively using EMO-DB dataset, 9.16%, 4.94%, and 11.68% for general, male, and female test using RAVDESS dataset and, 6.91% for general test using SAVEE dataset on speaker-dependent experiments is obtained.

The detailed results which also include emotion-wise recall and accuracy for both speaker-dependent and speaker-independent scenarios are provided in section 3.

The rest of the paper is organized as follows. Explanation of techniques used, emotional datasets, proposed feature optimization algorithm and feature selection methods are presented in Section 2. Experimental results are given in Section 3. Section 4 concludes the paper.

II. MATERIALS AND METHODS

In this section, the theoretical underpinning of the techniques utilized in this research will be established. Firstly, the concept of well-known techniques, GA and DBSCAN will be revisited. Then the proposed feature optimization and selection technique based on GA and DBSCAN will be elaborated. Lastly, the concepts of classification algorithm (SVM) used in this study, will be discussed in detail.

A. GA

The GA [30] has taken its analogy from the genetic process of living organisms. The solution space of GA consists of chromosomes or individuals. A chromosome has the genetic information for each individual. A set of chromosomes make a population. The priority of each chromosome is evaluated by using a fitness function. Chromosomes which are considered fit are selected for recombination through a crossover step to produce a new individual or offspring. After that mutation is applied on population to introduce the randomness. The detail on the GA is provided in the following section.

1) POPULATION

A set of possible solutions are termed as population. The best possible solutions are found by applying GA operators. The population size can be kept fixed or variable throughout the optimization process [31]. For longer chromosome, the size of population should be big [32], however it needs more iterations and results in consuming more computational time with slow convergence.

2) CROSSOVER

Crossover starts with selecting parents, usually by random selection. After selecting mating population, the crossover operator Pc is applied. It determines if the current parent chromosome is crossed over or if it is moved directly to the offspring population [32]. The crossing over can be applied as uniform or non-uniform manner. Uniform operators act in a similar way in every generation. Whereas non-uniform operators work according to the age of population. The mathematical representation of two offspring formed as a result of crossover of two parents is as follows:

$$y_i^1 = \alpha_i X_i^1 + (1 - \alpha_i) X_i^2 \quad (1)$$

$$y_i^1 = \alpha_i X_i^2 + (1 - \alpha_i) X_i^1 \quad (2)$$

where α_i stands for uniform random number. α will be constant in uniform crossover and variable in non-uniform crossover.

3) MUTATION

The mutation operator is responsible for random changes in the population. Like natural biological process, the diversity in the population is maintained by mutation. It also has uniform and non-uniform operators. The mutation probability is denoted by P_m. Too big value of P_m could result in loss of good genetic material, similarly, too small of a value would keep the population the same.

B. DBSCAN

DBSCAN (Density Based Spatial Clustering of Applications with Noise), devised by Ester et. al [33] is a density based clustering technique. It can learn the clusters of arbitrary shapes. Initially the algorithm was proposed for clustering spatial data, however, due to its unique features, it has been applied for clustering other data types in numerous fields. In civil engineering it is used for clustering spatial civil infrastructure networks [34], in spectroscopy it is used for grouping single particle mass spectra [35], in medical science it is used for lesion detection [36]. Additionally, It is fast when clustering small to medium size data sets [37].

The algorithm takes three parameters. The first parameter is the matrix of $m \times n$ dimensions where m is the number of objects or observations and n is the number of features for each observation. The second parameter is *Eps* which is neighborhood size, and *Minpts* is number of objects in that neighborhood. The clustering scheme identifies three different types of objects, namely, core objects, border objects and noise. The points which are neither border, nor core are noise. The mathematical definitions of these concepts [33] are provided here and Figure 1 clarifies the above points in a visual way. The nice thing about DBSCAN is that one does not need to specify the number of clusters to use it. In best case the time complexity of DBSCAN is $O(nlogn)$, however the worst case time complexity is $O(n^2)$.

Definition 1 (Eps): The *Eps* of a point p is denoted as Eps_p and defined as:

$$Eps_p = q \in D | dist(p, q) <= Eps.$$

Definition 2 (Cluster): Let D be a dataset of points. A cluster $C$is a non-empty subset of D wrt. *Eps* and *Minpts* having satisfied the following:

1) $\forall U, V$: if $U \in C$ *and* V is density-reachable from U wrt. Eps and Minpts then $V \in C$.

Definition 3 (Core Object): An object or observation is considered as core if

$$density(coreObject) <= Minpts$$

Definition 4 (Border Object): The object is border if

$$density(border) < Minpts$$
$$ANDborderObject \in neighborhood(core)$$

Definition 5 (Noise): Let $C_1, C_2, \ldots\ldots C_n$, be the clusters of dataset D wrt. eps_i and $Minpts_i$ where $i = 1, 2, \ldots..n$. The noise or outlier is defined as a set of points in the dataset D not belong to any cluster C_i

$$\text{noise} = u \in D | \forall i : u \notin C_i$$

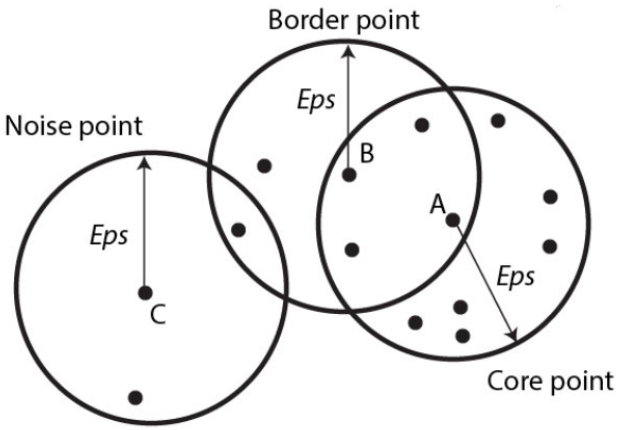

FIGURE 1. Visual representation of DBSCAN parameters.

C. DESCRIPTION OF DATA SETS

1) EMO-DB

The EMO-DB is an acted dataset of ten professionals (five male and five female) recorded by F. Burkhardt in the German language [38]. It is labeled with seven emotion classes of anger, boredom, fear, happy, disgust, neutral, and sadness. There are multiple utterances of the same speaker. Ten sentences, which are linguistically neutral, are chosen for dataset construction. Out of these 10 sentences, 5 sentences are short (approximately 1.5 Sec long) and 5 are long sentences (approximately 4 Sec long). Each emotion class has nearly equal number of emotional utterances in order to avoid the problem of under-sampling emotion class. There are a total of 535 utterances in this dataset. It is one of the most widely used datasets in the literature [26]. The dataset includes those utterances which have a recognition rate of more than 80% in a subjective listening test.

2) RAVDESS

RAVDESS is an approved multi-modular database of emotional speech and song [39]. The dataset is multipurpose having the modalities of audio-visual, video-only and audio-only. Here in this research the audio-only modality is used for speech emotion recognition task. There are 24 professional actors each uttering 60 unique intonations for speech with emotions: happy, sad, angry, fear, surprise, disgust, calm, and neutral. The RAVDESS dataset is exceptionally rich in nature giving that it doesn't experience the gender bias, comprises of a wide range of emotions, and has two levels of emotional intensity. Each actor uses two different statements with intensities, normal and strong for each emotion except for neutral which is with the normal intensity only. Each unique recording is rated 10 times for emotional validity,

intensity and originality. The total number of utterances is (60 * 24) = 1440.

3) SAVEE

SAVEE dataset provides audio utterances of British speakers [40]. The speakers were four British male actors who spoke the sentences showing six emotions: anger, sadness, disgust, happiness, surprise, fear. The sentences chosen were phonetically balanced. The dataset comprises of 480 audio utterances. The data is processed and labeled under the visual media lab using high quality audio and video equipment.

D. FEATURE EXTRACTION

This stage involves extraction of useful features for speech emotion analysis. Well known toolboxes for feature extraction in speech signals are OpenSMILE [41], OpenEAR [42], HTK [43], and Praat [44]. In this research, OpenSMILE tool box is used to extract INTERSPEECH 2010 Challenge feature set consisting of 1582 features. [45]. The reason for opting INTERSPEECH 2010 feature set is its coverage for broad categories of features namely: prosodic, spectral, and energy effective for emotion recognition. This fact was proved in [46] where it gave the best results for emotion recognition with many classifiers. The summary of features is provided in Table 1.

TABLE 1. Summary of Interspeech 2010 challenge feature set extracted using OpenSMILE Toolkit.

Interspeech 2010 challenge acoustic features	no. of features
MFCCs	630
Log power of mel-frequency bands	336
Pair frequencies of line spectrum	336
Loudness	42
Smoothed fundamental frequency contour's envelope	42
Final fundamental frequency voicing probability	42
Contour of smoothed fundamental frequency	40
Local shimmer	38
Local jitter	38
Differential jitter	38
Total	1582

E. FEATURE OPTIMIZATION USING PROPOSED METHOD

This section will briefly describe the proposed DGA process. To compare the performance of proposed optimization algorithm, the PCA only and standard GA with PCA will be used. The proposed method is different to the standard GA with respect to selection process. Instead of random selection of parents for crossover and mutation, we are applying density based clustering. The similar data objects are placed in the same cluster, while objects having low similarity with any group are considered outliers. This serves two purposes. Firstly, the population for crossover and mutation is selected only from clusters (non-outliers). Secondly, the newly generated population replaces outliers thus ensuring sufficient pressure to obtain even better population from current individuals. The DGA is described in Algorithm 1.

Firstly, the database of normalized features is loaded and emotion classes are separated. The initial population is selected from the single emotion class. After that DBSCAN clustering is applied to detect outlier. In the next step, parents for cross over and mutation are selected from non-outliers. For cross over it is ensured that both parents belong to the same gender so that gender specific emotion information do not get mixed up. The next step allows single point cross over with probability 0.9 and mutation with probability 0.1. In the next step the outliers detected earlier are replaced with newly created children having the same gender. The whole process is repeated until the stopping criteria is met and all emotion classes are dealt.The flow chart of the proposed technique is given in Figure 2. The stopping criteria for optimization process is kept in double the size of dataset because we are able to obtain maximum results on it. At every iteration, new outliers are detected and non-outliers are selected for participation in crossover and mutation. The resulting feature set is optimized, however, we need to reduce its dimensions

F. FEATURE REDUCTION

The feature set is optimized however, huge, consisting of 1582 features, which require a good dimensionality reduction technique. Among the effective methods available(Forward Feature Selection (FFS), Backward Feature Selection (BFS) [47], Principle Component Analysis (PCA), and Linear Discriminate Analysis (LDA) [40], [48]), we opted for the most commonly used PCA [49]–[51]. PCA includes finding the eigenvalues and eigenvectors of the available covariance matrix, and choosing the necessary number of eigenvectors compared to the biggest eigenvalues to create a transformed matrix. The matrix is utilized to change the original feature set into a transformed feature space and select the best-required features. After going through the optimization process by our proposed DGA, We fed the openSMILE INTERSPEECH 2010 feature set to PCA. The reduced feature set consists of 100 features that have further been used in the classification step.

G. SUPPORT VECTOR MACHINE

SVM is superior to existing methods because of its structural risk minimization approach and thus having better discrimination power [52]. It has the ability to solve non-linearly separable problems by the use of kernel functions. The SVM classifier endeavors to isolate samples belonging to two considered classes by maximizing the margins of hyperplane in the original feature space or in a high dimensional feature space by nonlinear mapping function $\varphi(.)$. In both cases the learning of the SVM depends on the mix of two criteria: i) empirical error minimization, and ii) control of model complexity. The former goes for optimizing the classification results in terms of accuracy on the training samples; the later controls the limit of the capacity utilized for abstaining from

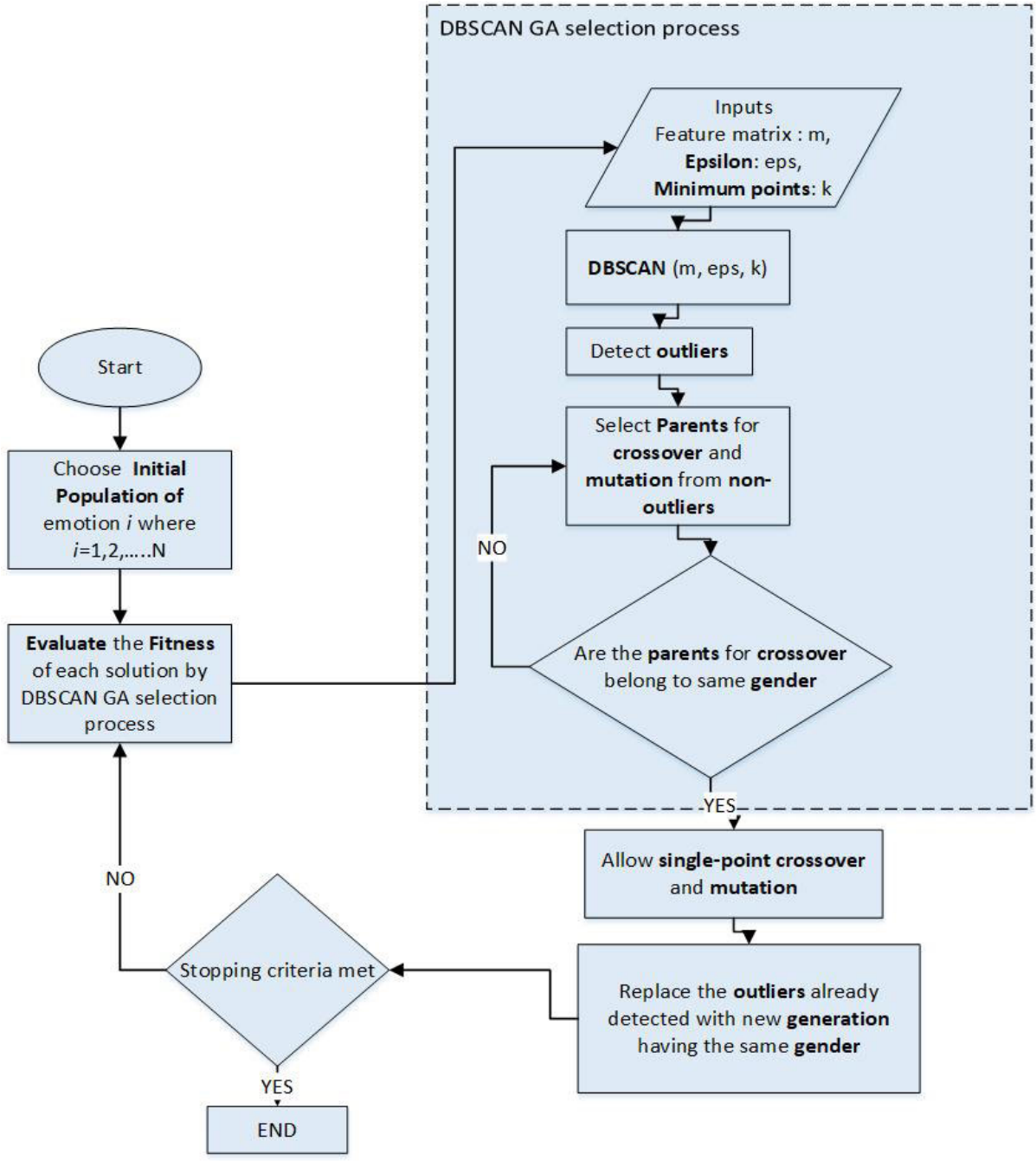

FIGURE 2. The flow chart representation of proposed DGA technique.

over fitting. These criteria are consolidated for defining the cost function to be minimized.

In case of linear SVM the cost function can be defined as

$$f(x) = \Sigma_{i=1}^{m} w\langle x, x_i\rangle + b \tag{3}$$

where w is a vector normal to hyper plane and b is a constant s.t $\frac{b}{\|w\|^2}$ represents the distance of the hyper plane to the origin. The classical SVM can be seen in Figure 6.

If the data in the feature space cannot be linearly separated, they can be anticipated into a higher dimensional feature space with a nonlinear mapping function $\varphi(.)$ characterized as per the Cover's theorem [53]. As a result the inner product between two mapped feature vectors becomes

$$f(x) = \Sigma_{i=1}^{m} w\langle \varphi(x), \varphi(x_i)\rangle + b \tag{4}$$

The function $f(x)$ can be derived by minimizing the following cost function, which is a tradeoff between empirical error minimization and solution complexity:

$$\frac{1}{2} \parallel w \parallel^2 + C\Sigma_{i=1}^{m} \begin{cases} [w.\varphi(x_i) + b] \geq 1 - \xi_i, \\ \quad i = 1, 2, \ldots., m \\ \xi_i \geq 0 \; and \; C < 0 \end{cases} \tag{5}$$

Algorithm 1 Clustering Based GA for Feature Optimization

Input: $S = (x_1, x_2, x_3, \ldots\ldots x_n, y)$: Dataset
$m = Chromosome - length$ // feature set size
Pop = Population
Pop_{size} =Population size
P_c =Cross over probability
P_m = Mutation probability
N =Number of emotion classes
Output: Best Fitness and Optimal Feature set
Initilize algorithm parameters
Load normalized dataset S
Separate each emotion class N from dataset S
for $i = 1$ *to* M **do**
  for $j = 1$ *to* Pop_{size} *of* N_i **do**
    $P_c = 0.9$ //initilize Pc
    $P_m = 0.1$ // initilize Pm
    $NewPop_{size} = Pop_{size}$
    $Eps = (0.01 - 0.1) *$ breath of N_i// adjust the Eps for DBSCAN
    $K = log(Chromosome - length)$// calculate the parameter K for DBSCAN
    Apply DBSCAN on *pop*
    Detect outliers
    select parent1 and parent2 from non-outliers and having the same gender
    if $P_c < 0.9$
      do Cross over
      replace the outliers with newly generated children from *pop* generation
    select parent3 from non-outliers
    if $P_m < 0.1$
      do Mutation
      replace the outliers with newly generated children
  end for
$New_{pop} = pop$ // update the existing population
end for
return optimized feature set of dataset S

where C is a regularization parameter, ξ_i are non-negative slack variables for dealing with noise and nonlinearly separable data, ω_i is the label of training set x_i, and m is the total number of training samples. The final decision function can be written as:

$$\omega = sign[f(x)] \tag{6}$$

The minimization problem in equation 3 can be solved according to the Lagrange theory obtaining a dual problem in which the following convex objection function should be maximized:

$$w(a) = \Sigma_{i=1}^{m}\alpha_i - \frac{1}{2}\begin{cases}\Sigma_{i=1}^{m}\omega_i\alpha_i = 0\\ 0 \le \alpha_i \le C \text{ and } C > 0\\ \quad i = 1, 2, \ldots, N\end{cases} \tag{7}$$

The Lagrangian $w(\alpha)$ should be maximized with respect to Lagrange multipliers α_i which are associated with training points x_i. Points associated with nonzero lagrange multipliers are called support vectors: The ones belonging to $0 < \alpha_i < C$ are called non-bound support vectors and lie inside the margin, while the ones belonging to $\alpha_i = C$ are called bound support vectors and lie on margin. These examples can be viewed as erros since they are related to a nonzero ξ_i. Support vectors are the main examples in the training set that decide the ideal hyperplane position. In non-linear SVM the function $\xi_.$ is unknown yet and QP problem solution is not possible using equation 5. According to Mercer's theorem, the inner product will be replaced by positive defined kernel function K(.). Now it is possible to avoid the inner details of feature vector i.e.

$$\langle \xi_{xi} \rangle = K(x_i, x_j)$$

Accordingly it is possible to prove that the discriment function can be rewritten in dual formulation

$$f(x) = \Sigma_{i=1}^{m}\alpha_i\omega_i K(x, x_i) + b \tag{8}$$

where b is calulated using the primal-dual relationship [54], and the samples which affect the solution are nonzero lagrange multipliers α_i. Thus the decision function is formed by applying (4) to (7). The kernels which satisfy Mercer's conditions and widely used are following.

Linear kernel: $k(x_i, x_j) = x_i.x_j$

Polynomial kernel: $k(x_i, x_j) = (1 + x_i.x_j)^d, d \in R^+$

Gaussian kernel: $k(x_i, x_j) = exp(- \parallel x_i - x_j \parallel^2 /2\sigma^2)$, $\sigma \in R^+$

where d and σ in polynomial and Gaussian kernels are adjustable parameters respectively. The Mercer's kernel ensures that there is no local maxima in the function to be optimized [55]. The classical SVM is defined as binary classifier, which can discriminate between two classes. However, there are several techniques in literature which make it possible to use SVM for multiclass classification problem with same discriminating power. Among the others are: The One-Againt-All(OAA) and One-Against-One(OAO) strategies [56]. Let $\Omega = \omega_1, \ldots, \omega_R$ be R different classes to be identified. In case of OAA strategy, R different binary SVMs are trained.Every binary classifier has to recognize the samples of a generic class $\omega \in \Omega$ from the samples of all the rest of the classes $\Omega - \omega_i$. A given pattern is marked by the class of the classifier that outcomes highest output value. In OAO architecture, for each pair of classes ω_i and ω_j where $i \neq j$, one classifier is considered. As a whole, we have $R(R-1)/2$ classifiers. Simple majority voting algorithm is used for classification.

H. EXPERIMENTAION

All the experiments and code implementations were carried out on MATLAB and they were executed on Desktop PC with 3.20 GHz Dual core i7 8700 processor and 32 GB RAM. For classification SVM was used, for which MATLAB built-in function, i.e fitcsvm with Radial Basis Function (RBF)

as kernel function and Sequential Risk Minimization (SMO) method for parameter optimization was employed. For faster convergence of SVM, the training samples were scaled to [0,1] using z-score normalization. The minimum and maximum values found during the scaling of training data, were also used to scale the test data as well.

TABLE 2. Parameter setting of feature optimization and selection.

Method	Parameters
DGA	Popsize = size of emotion class, iter = double the size of pop K = log(feature set size), Eps = adjusted according to dataset Pm = 0.1, Pc = 0.9, size of dimension = 1582
PCA	Initial feature set size = 1582 Final feature set size = 100

Out of the two commonly used approaches to construct multi-class SVM classifier: 1) One-Against-All (OAA) and 2) One-Against-One (OAO), OAA is one of the most widely used method [57]. It is pairwise classification where there is one binary SVM for each pair of classes to discriminate one class from the rest of the classes. As compared to OAA, OAO constructs one SVM to distinguish each pair of classes. Here we are using the OAA approach for the sake of better recognition accuracy. In order to classify only one emotion at a time, we are choosing the class with largest interval value [58].

To evaluate the performance of DGA along with PCA, we need to define some parameters. The list of these parameters is given in Table 2. The parameter *popsize* is the size of population which is equal to the size of emotion class as given in Table 3 and *iter* is the number of iterations needed for optimization. *Pm* is the probability for mutation and *Pc* is the probability for crossover. In case of clustering, *Eps* is a neighborhood size and *K* is the number of objects in that neighborhood.

We have mainly two scenarios for experimentation: Speaker-dependent and speaker-independent. For speaker-dependent experiments 7-folds cross validation is performed. In literature 10-folds cross validation is commonly used. However, in our case, we have experimented with 10-folds and came to know that the size of validation set became so small that some of the samples were totally missed out from test set. So, keeping in view the datasets we are using in this study, by using the 7-folds we do not have risk of insufficient samples per emotion class per speaker to validate the results.In case of speaker-independent experiments, leave-one-subject-out (LOSO) test is performed. LOSO requires model to be trained with $1..(n-1)$ speakers and tested with *nth* speaker. The process is repeated for each speaker. Speaker-dependent experiments cover general (combining all speakers), male only and female only tests whereas speaker-independent experiments include only general tests. For feature selection, PCA is employed only on training data and using the training-data-coefficients of PCA, the dimensions of testing data are reduced.

TABLE 3. Size of each emotion class.

Dataset	Emotion	Emotion class size
EMO-DB	Disgust	44
	Anger	127
	Happy	71
	Fear	66
	Sadness	62
	Neutral	78
	Boredom	80
RAVDESS	Neutral	88
	Calm	172
	Happy	187
	Sadness	183
	Anger	191
	Fear	191
	Disgust	181
	Surprise	182
SAVEE	Anger	60
	Sadness	60
	Surprise	60
	Disgust	60
	Fear	60
	Happy	60
	Neutral	60

The metrics for results calculation are unweighted recall (UAR) and accuracy. Recall is calculated for each individual class. Recognition accuracy is measured by averaging the individual class recognition rate weighted by the priority of class.

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{9}$$

$$Weight_i = \frac{TP_i + FN_i}{N} \tag{10}$$

$$Accuracy = \sum_{i=1}^{M} WeightXRecall_i \tag{11}$$

$$UAR = \frac{1}{M} \sum_{i=1}^{M} Recall_i \tag{12}$$

In case of the 7-fold cross validation the accuracy and recall of each individual fold is averaged to get the average accuracy and recall. In speech emotion recognition research, accuracy and UAR are standard evaluation measures as number of utterances in each emotion class are different [59]. The entire research methodology covering preprocessing, feature optimization, feature reduction and model representation is pictorially shown in Figure 3.

III. RESULTS

The results are tabulated according to speaker-dependent and speaker-independent scenarios.

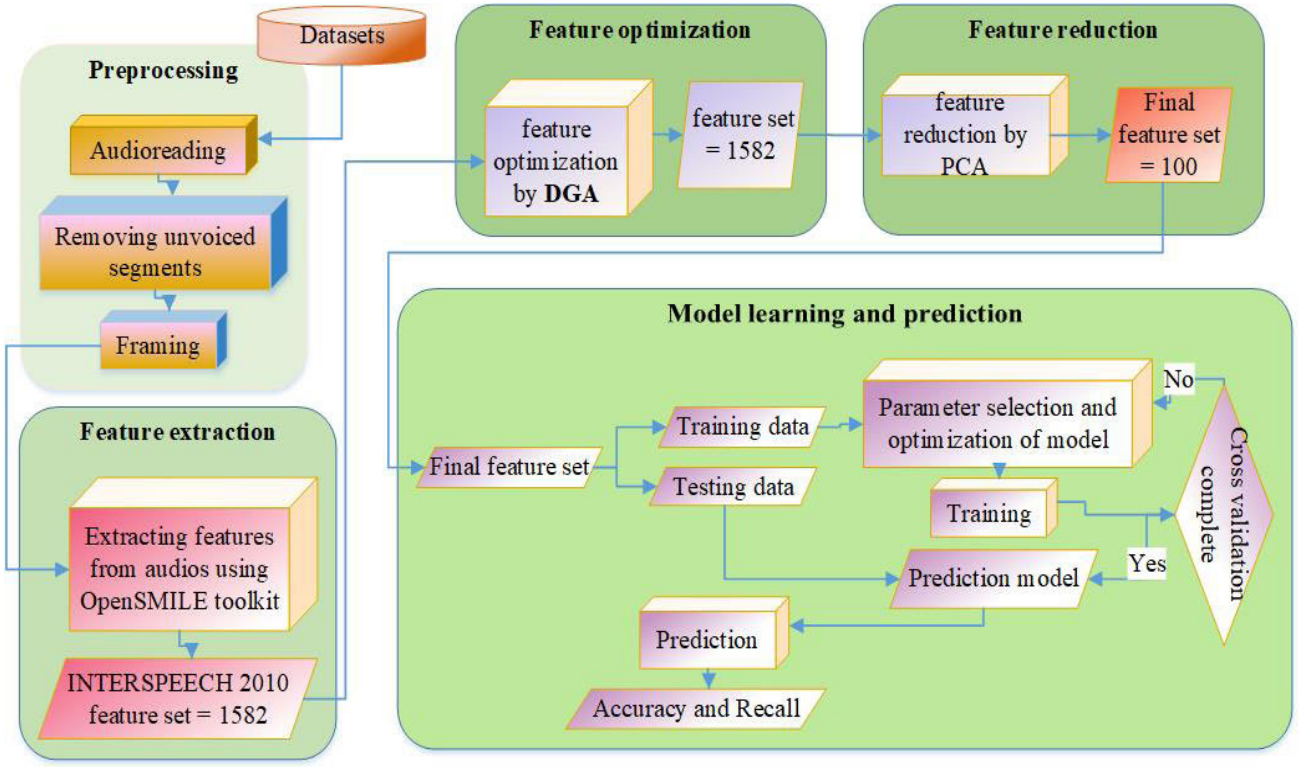

FIGURE 3. Research methodology.

TABLE 4. 7-fold cross-validation accuracy results for speaker-dependent experiments.

Dataset	Classifier		Feature optimization and Feature selection PCA only	Standard GA + PCA	DGA + PCA
EMO-DB	SVM				
		general	84.09	87.7605	**89.6589**
		male	66.77	71.3991	**86.2648**
		female	84.51	87.0425	**88.3875**
RAVDESS	SVM				
		general	72.66	73.43	**82.5907**
		male	70.94	70.55	**75.497**
		female	78.63	79.44	**91.1207**
SAVEE	SVM				
		general	63.45	70.83	**77.74**

Firstly speaker-dependent evaluation will be done and then speaker-independent evaluation.

A. SPEAKER-DEPENDENT EXPERIMENTS

This section explores the recognition of speech emotions for the speaker-dependent scenario. Table 4 shows the recognition results for EMO-DB dataset for speaker-dependent case.

Feature optimization and feature selection (DGA + PCA) has improved results from 84.09% to 87.76% for general test as compared to when only PCA is used. The standard GA gained its accuracy on gender-dependent tests rising from 66.77% to 71.39% and 84.81% to 87.04 % for male and female speakers respectively as compared to PCA. In comparison to this, the highest accuracy was consistently observed for all three experiments for general, male and female tests as 89.66%, 86.26%, and 88.38% respectively using the proposed DGA + PCA method.

For RAVDESS dataset, the current human accuracy is reported as 67% [39] which reflects that the recognition of emotions for this dataset was not an easy and straightforward task even for human beings. Our algorithm (DGA + PCA) not only outperformed the standard GA + PCA but remarkably improved the recognition rate as 82.59% for general test, 75.49% for male tests and, 91.12% for female tests.

The SAVEE dataset is consists of male speakers only, so gender-based results cannot be calculated and compared. The general tests result gave the highest accuracy gain of 77.74% by using our proposed algorithm

TABLE 5. Recall values (%) of each emotion class: speaker-dependent experiments.

Dataset	Emotion	Feature optimization and Feature selection PCA only	Standard GA + PCA	DGA + PCA
EMO-DB	Disgust	77.55	88.77	**90.81**
	Anger	91.35	93.73	**95.27**
	Happy	64.93	71.81	**85.84**
	Fear	86.34	**87.77**	84.44
	Sadness	**96.62**	95.04	93.65
	Neutral	**86.03**	84.74	84.52
	Boredom	86.36	92.42	**92.42**
RAVDESS	Neutral	41.20	44.59	**75.18**
	Calm	81.90	82.47	**90.71**
	Happy	67.88	68.94	**76.55**
	Sadness	65.50	64.95	**74.94**
	Anger	84.75	85.82	**86.86**
	Fear	75.43	74.92	**83.76**
	Disgust	82.92	82.92	**88.46**
	Surprise	81.86	82.96	**84.61**
SAVEE	Anger	69.64	82.14	**83.53**
	Sadness	60.31	64.48	**75.96**
	Surprise	61.70	**72.81**	70.04
	Disgust	52.18	59.92	**65.67**
	Fear	50.59	51.78	**78.57**
	Happy	58.53	68.45	**78.57**
	Neutral	92.57	**95.84**	92.57

TABLE 6. 7-fold cross-validation accuracy results for speaker-independent experiments.

Dataset	Classifier		Feature optimization and Feature selection PCA only	Standard GA + PCA	DGA + PCA
EMO-DB	SVM	general	60.06	69.77	**77.49**
RAVDESS	SVM	general	53.26	60.11	**76.20**
SAVEE	SVM	general	29.76	46.07	**69.88**

as compared to the state-of-the-art PCA only and GA + PCA which has 63.45% and 70.83% accuracy respectively.

The effectiveness of proposed optimization method on individual emotion classes are shown in Table 5 on EMO-DB, RAVDESS, and SAVEE datasets for speaker-dependent experiments. We have taken UAR measure for each emotion class. In case of EMO-DB dataset, the recall rate of the proposed method is highest for four out of seven emotion classes with the performance gain of 2% for disgust, 1.54% for anger, and 14.03% for happy with respect to standard GA + PCA. Whereas, in the case of RAVDESS dataset it is giving the maximum score of recall for all eight emotion classes having the highest performance gain of 30% for neutral utterances when compared with standard GA + PCA. The performance gain was 8.24% for calm, 7.61% for happy, 10% for sadness, 1.04% for anger and 8.84% for fear. For SAVEE dataset the proposed method outperformed for five out of seven emotion classes with the gain of 1.39% for anger, 11.48% for sadness, 5.75% for disgust, 26.79% for fear, and 10.12% for happy. These results on three datasets are showing different rate of performance gain for each emotion class. For example, in case of EMO-DB, the emotion happy is showing the highest performance gain, whereas in case of RAVDESS, the neutral state is showing highest recognition rate and performance gain. In case of SAVEE dataset, it is the emotion of fear which is showing highest performance gain. Because the datasets are imbalanced and have different number of emotion classes, we cannot expect the same trend of performance gain and also we can't infer any reason for the variation in performance gain. However, it is quite evident, that our proposed method improved the recognition rate for almost all of the emotion classes.

These results as visually represented in Figures 4, 5 and, 6 are highlighting the performance of our proposed method in comparison to the baseline methods.

TABLE 7. Recall values (%) of each emotion class: speaker-independent experiments.

Dataset	Emotion	Feature optimization and Feature selection PCA only	Standard GA + PCA	DGA + PCA
EMO-DB	Disgust	55.15	69.74	**84.02**
	Anger	68.95	72.87	**95.37**
	Happy	47.94	41.59	**63.45**
	Fear	49.82	52.56	**78.81**
	Sadness	70.14	**88.48**	80.83
	Neutral	39.32	59.24	**62.93**
	Boredom	48.29	64.65	**71.25**
RAVDESS	Neutral	41.20	44.59	**75.18**
	Calm	81.90	82.47	**90.71**
	Happy	67.89	68.94	**76.56**
	Sadness	65.51	64.96	**74.95**
	Anger	84.75	85.83	**86.87**
	Fear	75.43	74.92	**83.77**
	Disgust	82.92	82.92	**88.46**
	Surprise	81.87	82.96	**84.61**
SAVEE	Anger	66.66	58.33	**73.33**
	Sadness	16.66	33.33	**65.00**
	Surprise	13.33	53.33	**65.00**
	Disgust	25.00	40.44	**53.33**
	Fear	20.00	41.66	**81.66**
	Happy	21.66	33.33	**71.66**
	Neutral	40.00	59.16	**79.16**

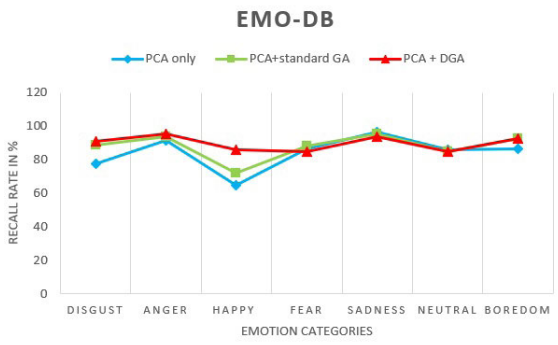

FIGURE 4. Emotion-wise average recall rate for speaker-dependent scenario using EMO-DB dataset.

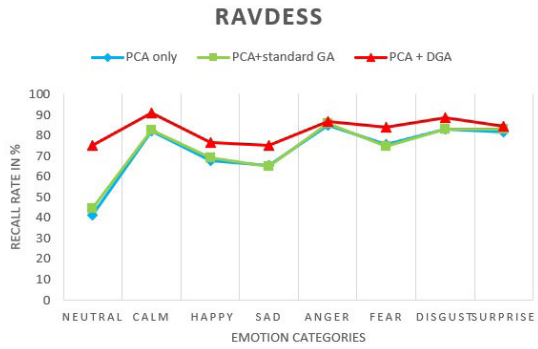

FIGURE 5. Emotion-wise average recall rate for speaker-dependent scenario using RAVDESS dataset.

B. SPEAKER-INDEPENDENT EXPERIMENTS

Speaker-independent case is explored in this section for speech emotion recognition. The accuracy measures on EMO-DB, RAVDESS and SAVEE datasets for only general tests are given in 6. The proposed method (DGA + PCA) has outperformed the state-of-the-art methods of PCA only and, Standard GA + PCA with the ultimate accuracy of 77.49% for EMO-DB, 76.20% for RAVDESS and, 69.88% for SAVEE dataset.

The effects of feature optimization and selection on individual emotion classes are given in Table 7 on all three datasets for speaker-independent experiments. The recall measure for each emotion class was calculated. In case of EMO-DB dataset, the recall rate of the proposed method is highest for six out of seven emotion classes, whereas in case of RAVDESS dataset it is giving the maximum score of recall for all eight emotion classes. In case of SAVEE dataset again the proposed method outperformed for seven out of seven emotion classes. These results give one important insight about the proposed method that it was good at discriminating individual emotions for speaker-independent experiments as compared to speaker-dependent experiments.

C. COMPARISON WITH EXISTING WORK IN THE LITERATURE

The results presented in this paper with respect to EMO-DB, RAVDESS, and SAVEE datasets can be compared with results of a few benchmark studies (performed on same

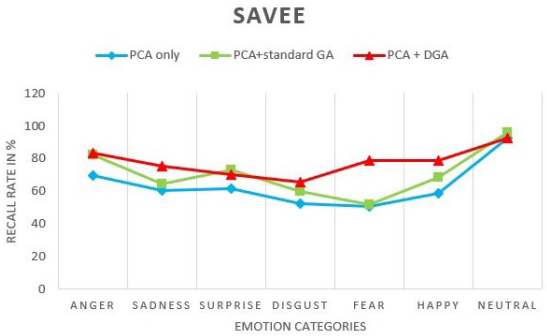

FIGURE 6. Emotion-wise average recall rate for speaker-dependent scenario using SAVEE dataset.

TABLE 8. Comparison of recognition accuracy with previous works using EMO-DB dataset.

Method	Speaker-dependent	Speaker-independent
[60]	85.57 %	-
[3]	85.82%	-
[59]	87.66%	-
[61]	-	71.7%
[62]	-	52%
[12]	77.67%	68.89%
[63]	86.96%	77.08%
[64]	87.8 %	-
[65]	81.3 %	-
DGA(proposed)	**89.65 %**	**77.49%**

datasets) available in literature. Table 8 compares the proposed results on speaker-dependent and speaker-independent experiments (corresponding to EMO-DB dataset) of this study with the results presented in [3], [12], [59]–[65]. It is evident from the Table 8 that the clustering-based GA improve the recognition performance of the SER systems when compared with existing work.

Table 9 compares the proposed results on speaker-dependent and speaker-independent experiments (corresponding to RAVDESS dataset) of this study with the results presented in [7], [60], [64], [66]–[68]. It is evident from Table 9 that the proposed feature optimization algorithm improves the recognition accuracy of the SER when compared to the existing systems.

Table 10 compares the obtained results (corresponding to SAVEE dataset) of this research on speaker-dependent and speaker-independent experiments with the results presented in [3], [19], [63], [69], [70]. It is clear from Table 10 that our proposed clustering base GA improve the recognition performance of the SER system when compared to the state of the art systems.

IV. DISCUSSION AND CONCLUSION

SER is a complex task, with two main stages, feature engineering and classification. In this study, we propose a new feature optimization algorithm used in combination with PCA on INTESPEECH 2010 feature set. Our proposed algorithm outperformed the baseline techniques for

TABLE 9. Comparison of recognition accuracy with previous works using RAVDESS dataset.

Method	Speaker-dependent	Speaker-independent
[60]	82.01%	-
[66]	81.3%	73.5%
[7]	-	71.6%
[67]	77.8%	-
[68]	79.5%	
[64]	82.3 %	-
DGA(proposed)	**82.5%**	**76.2%**

TABLE 10. Comparison of recognition accuracy with previous works using SAVEE dataset.

Method	Speaker-dependent	Speaker-independent
[69]	48.41 %	-
[70]	75.6%	50.0%
[19]	76.19%	-
[3]	76.4%	44.18%
[63]	77.08%	55.83%
DGA(proposed)	**77.74%**	**69.88%**

RAVDESS, EMO-DB, and SAVEE datasets. The SER model based on the proposed feature optimization technique is also compared with many state-of-the-art studies in terms of accuracy and recall. It is evident from the comparison tables: Table 8-10, that the SER model based on the proposed optimization technique achieves comparable classification performance. Specifically, in the speaker-dependent experiments, the recognition rate of 89.65%, 82.5%, and 77.74% are obtained for EMO-DB, RAVDESS, and SAVEE datasets respectively. For speaker-independent experiments, we achieve the recognition rate of 77.49%, 76.2%, and 69.88% for EMO-DB, RAVDESS, and SAVEE datasets respectively. There are some other studies in the literature such as [71], [72], which are showing even better scores than ours. However, their parameters in terms of classification algorithms, feature sets, and feature engineering methods are different. For example, in [72], the recognition accuracy of 79.2% is achieved on EMO-DB dataset for the speaker-independent scenario, which is higher than ours. However, the focus of the study was feature extraction, whereas, in our study, feature optimization was the main concern.

In another study [71], the authors have proposed a meta-heuristic based feature selection method and achieved even higher scores on EMO-DB and SAVEE as 98.46% and 97.31% respectively on speaker-dependent experiments. The direct comparison even in this case is not possible as the feature set and classifier used in the study were different.

Nevertheless, more experiments can be done to achieve better accuracy. The inclusion of other feature types such as Teager energy operator or use of deep neural networks for classification could significantly improve the accuracy of the SER model. In addition, the use of auto-encoders and other feature selection methods can also improve the accuracy.

As future work, we will extend our experiments on the proposed method with more datasets, different combination of features, and with deep learning and other machine learning techniques in SER. Performing the time complexity analysis of the proposed method is also one of our plans.

REFERENCES

[1] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020.

[2] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, 2018.

[3] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, Oct. 2018.

[4] P. Song and W. Zheng, "Feature selection based transfer subspace learning for speech emotion recognition," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 373–382, Jul. 2020.

[5] A. Bhavan, P. Chauhan, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowl.-Based Syst.*, vol. 184, Nov. 2019, Art. no. 104886.

[6] A. R. Avila, Z. Akhtar, J. F. Santos, D. OShaughnessy, and T. H. Falk, "Feature pooling of modulation spectrum features for improved speech emotion recognition in the wild," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 177–188, Jan. 2021.

[7] D. Issa, M. F. Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101894.

[8] S. Chattopadhyay, A. Dey, and H. Basak, "Optimizing speech emotion recognition using manta-ray based feature selection," 2020, *arXiv:2009.08909*. [Online]. Available: http://arxiv.org/abs/2009.08909

[9] S. Wang, P. Phillips, Z.-C. Dong, and Y.-D. Zhang, "Intelligent facial emotion recognition based on stationary wavelet entropy and Jaya algorithm," *Neurocomputing*, vol. 272, pp. 668–676, Jan. 2018.

[10] R. V. Darekar and A. P. Dhande, "Emotion recognition from Marathi speech database using adaptive artificial neural network," *Biologically Inspired Cognit. Archit.*, vol. 23, pp. 35–42, Jan. 2018.

[11] D. Gharavian, M. Bejani, and M. Sheikhan, "Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2331–2352, Jan. 2017.

[12] F. Daneshfar and S. J. Kabudian, "Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm," *Multimedia Tools Appl.*, vol. 79, nos. 1–2, pp. 1261–1289, Jan. 2020.

[13] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Feature extraction based on bio-inspired model for robust emotion recognition," *Soft Comput.*, vol. 21, no. 17, pp. 5145–5158, Sep. 2017.

[14] A. Chehouri, R. Younes, J. Khoder, J. Perron, and A. Ilinca, "A selection process for genetic algorithm using clustering analysis," *Algorithms*, vol. 10, no. 4, p. 123, Nov. 2017.

[15] C. Yogesh, M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, C. Berkai, and K. Polat, "A new hybrid PSO assisted biogeography-based optimization for emotion and stress recognition from speech signal," *Expert Syst. Appl.*, vol. 69, pp. 149–158, Mar. 2017.

[16] A. M. Badshah, B. N. Rahim, N. Ullah, J. Ahmad, K. Muhammad, M. Y. Lee, S. Kwon, and S. W. Baik, "Deep features-based speech emotion recognition for smart affective services," *Multimedia Tools Appl.*, vol. 78, no. 5, pp. 5571–5589, Mar. 2019.

[17] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[18] C. Brester, E. Semenkin, and M. Sidorov, "Multi-objective heuristic feature selection for speech-based multilingual emotion recognition," *J. Artif. Intell. Soft Comput. Res.*, vol. 6, no. 4, pp. 243–253, Oct. 2016.

[19] C. K. Yogesh, M. Hariharan, R. Ngadiran, A. H. Adom, S. Yaacob, and K. Polat, "Hybrid BBO_PSO and higher order spectral features for emotion and stress recognition from natural speech," *Appl. Soft Comput.*, vol. 56, pp. 217–232, Jul. 2017.

[20] Z. Peng, X. Li, Z. Zhu, M. Unoki, J. Dang, and M. Akagi, "Speech emotion recognition using 3D convolutions and attention-based sliding recurrent networks with auditory front-ends," *IEEE Access*, vol. 8, pp. 16560–16572, 2020.

[21] H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech emotion recognition from 3D Log-Mel spectrograms with deep learning network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019.

[22] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 143–160, Jun. 2013.

[23] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1162–1171, Nov. 2011.

[24] R. B. Lanjewar, S. Mathurkar, and N. Patel, "Implementation and comparison of speech emotion recognition system using Gaussian mixture model (GMM) and k-nearest neighbor (K-NN) techniques," *Proc. Comput. Sci.*, vol. 49, pp. 50–57, Aug. 2015.

[25] S. Prasomphan and S. Doungwichain, "Detecting human emotions in a large size of database by using ensemble classification model," *Mobile Netw. Appl.*, vol. 23, no. 4, pp. 1097–1102, Aug. 2018.

[26] M. B. Mustafa, M. A. M. Yusoof, Z. M. Don, and M. Malekzadeh, "Speech emotion recognition research: An analysis of research focus," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 137–156, Mar. 2018.

[27] H. He and Y. Tan, "Automatic pattern recognition of ECG signals using entropy-based adaptive dimensionality reduction and clustering," *Appl. Soft Comput.*, vol. 55, pp. 238–252, Jun. 2017.

[28] S. Borra, R. Thanki, and N. Dey, *Satellite Image Analysis: Clustering and Classification*. Singapore: Springer, 2019.

[29] M. C. Thomas, W. Zhu, and J. A. Romagnoli, "Data mining and clustering in chemical process databases for monitoring and knowledge discovery," *J. Process Control*, vol. 67, pp. 160–175, Jul. 2018.

[30] S. Sivanandam and S. Deepa, "Genetic algorithms," in *Introduction to Genetic Algorithms*. Berlin, Germany: Springer, 2008, pp. 15–37.

[31] D. E. Goldberg, "Sizing populations for serial and parallel genetic algorithms," in *Proc. 3rd Int. Conf. Genetic Algorithms*, 1989, pp. 70–79.

[32] D. E. Goldberg, "A note on Boltzmann tournament selection for genetic algorithms and population-oriented simulated annealing," *Complex Syst.*, vol. 4, no. 4, pp. 445–460, 1990.

[33] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, vol. 96, 1996, pp. 226–231.

[34] D. P. de Oliveira, J. H. Garrett, and L. Soibelman, "A density-based spatial clustering approach for defining local indicators of drinking water distribution pipe breakage," *Adv. Eng. Informat.*, vol. 25, no. 2, pp. 380–389, Apr. 2011.

[35] L. Zhou, P. K. Hopke, and P. Venkatachari, "Cluster analysis of single particle mass spectra measured at flushing, NY," *Analytica Chim. Acta*, vol. 555, no. 1, pp. 47–56, Jan. 2006.

[36] M. Mete, S. Kockara, and K. Aydin, "Fast density-based lesion detection in dermoscopy images," *Computerized Med. Imag. Graph.*, vol. 35, no. 2, pp. 128–136, Mar. 2011.

[37] T. N. Tran, K. Drab, and M. Daszykowski, "Revised DBSCAN algorithm to cluster data with dense adjacent clusters," *Chemometrics Intell. Lab. Syst.*, vol. 120, pp. 92–96, Jan. 2013.

[38] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, Sep. 2005, pp. 1–4.

[39] S. Livingstone and F. A. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLoS ONE*, vol. 13, no. 5, 2018, Art. no. e0196391, doi: 10.5281/zenodo.1188976.

[40] S. Haq, P. J. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP)*, 2008, pp. 1–6.

[41] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in *Proc. Int. Conf. Multimedia (MM)*, 2010, pp. 1459–1462.

[42] F. Eyben, M. Wollmer, and B. Schuller, "OpenEAR—Introducing the Munich open-source emotion and affect recognition toolkit," in *Proc. 3rd Int. Conf. Affect. Comput. Intell. Interact. Workshops*, Sep. 2009, pp. 1–6.

[43] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 429–442, Jul. 2000.

[44] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot Int.*, vol. 5, nos. 9–10, pp. 341–345, 2002.
[45] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The INTERSPEECH 2010 paralinguistic challenge," in *Proc. Interspeech*, Sep. 2010, pp. 1–4.
[46] T. Özseven and M. Düğenci, "SPeech ACoustic (SPAC): A novel tool for speech feature extraction and classification," *Appl. Acoust.*, vol. 136, pp. 1–8, Jul. 2018.
[47] T.-L. Pao, Y.-T. Chen, J.-H. Yeh, and Y.-H. Chang, "Emotion recognition and evaluation of Mandarin speech using weighted D-KNN classification," in *Proc. 17th Conf. Comput. Linguistics Speech Process.*, 2005, pp. 203–212.
[48] T. Özseven, M. Düğenci, and A. Durmuşoğlu, "A content analaysis of the research aapproaches in speech emotion recognition," *Int. J. Eng. Sci. Res. Technol.*, vol. 7, no. 1, pp. 1–26, 2018.
[49] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," *Digital Signal Process.*, vol. 22, pp. 1154–1160, Dec. 2012.
[50] K. R. Scherer, J. Sundberg, L. Tamarit, and G. L. Salomão, "Comparing the acoustic expression of emotion in the speaking and the singing voice," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 218–235, Jan. 2015.
[51] S. Patel, K. R. Scherer, E. Björkner, and J. Sundberg, "Mapping emotions into acoustic space: The role of voice production," *Biol. Psychol.*, vol. 87, no. 1, pp. 93–98, Apr. 2011.
[52] L. Cen, H. L. Z. L. Yu, M. Dong, and P. Chan, *Machine Learning Methods in the Application of Speech Emotion Recognition*. London, U.K.: INTECH Open Access Publisher, 2010.
[53] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. Electron. Comput.*, vol. EC-14, no. 3, pp. 326–334, Jun. 1965.
[54] B. Scholkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
[55] F. Bovolo, L. Bruzzone, and L. Carlin, "A novel technique for subpixel image classification based on support vector machine," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2983–2999, Nov. 2010.
[56] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Trans. Neural Netw.*, vol. 13, no. 2, pp. 415–425, Mar. 2002.
[57] N. Yang, J. Yuan, Y. Zhou, I. Demirkol, Z. Duan, W. Heinzelman, and M. Sturge-Apple, "Enhanced multiclass SVM with thresholding fusion for speech-based emotion classification," *Int. J. Speech Technol.*, vol. 20, no. 1, pp. 27–41, Mar. 2017.
[58] V. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.
[59] S. Yildirim, Y. Kaya, and F. Kılıç, "A modified feature selection method based on Metaheuristic algorithms for speech emotion recognition," *Appl. Acoust.*, vol. 173, Feb. 2021, Art. no. 107721.
[60] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020.
[61] J. Pittermann, A. Pittermann, and W. Minker, "Emotion recognition and adaptation in spoken dialogue systems," *Int. J. Speech Technol.*, vol. 13, pp. 49–60, Mar. 2010.
[62] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service (PlatCon)*, Feb. 2017, pp. 1–5.
[63] S. Nagarajan, S. S. S. Nettimi, L. S. Kumar, M. K. Nath, and A. Kanhe, "Speech emotion recognition using cepstral features extracted with novel triangular filter banks based on bark and ERB frequency scales," *Digit. Signal Process.*, vol. 104, Sep. 2020, Art. no. 102763.
[64] P. Nantasri, E. Phaisangittisagul, J. Karnjana, S. Boonkla, S. Keerativittayanun, A. Rugchatjaroen, S. Usanavasin, and T. Shinozaki, "A light-weight artificial neural network for speech emotion recognition using average values of MFCCs and their derivatives," in *Proc. 17th Int. Conf. Electr. Eng./Electron., Comput., Telecommun. Inf. Technol. (ECTI-CON)*, Jun. 2020, pp. 41–44.
[65] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Inf. Sci.*, vol. 509, pp. 150–163, Jan. 2020.
[66] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. B. Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors*, vol. 20, no. 21, p. 6008, Oct. 2020.
[67] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021.
[68] S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, vol. 20, no. 1, p. 183, Dec. 2019.
[69] M. Sidorov, C. Brester, W. Minker, and E. Semenkin, "Speech-based emotion recognition: Feature selection by self-adaptive multi-criteria genetic algorithm," in *Proc. LREC*, May 2014, pp. 3481–3485.
[70] Y. Sun, G. Wen, and J. Wang, "Weighted spectral features based on local Hu moments for speech emotion recognition," *Biomed. Signal Process. Control*, vol. 18, pp. 80–90, Apr. 2015.
[71] A. Dey, S. Chattopadhyay, P. K. Singh, A. Ahmadian, M. Ferrara, and R. Sarkar, "A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition," *IEEE Access*, vol. 8, pp. 200953–200970, 2020.
[72] K. Wang, G. Su, L. Liu, and S. Wang, "Wavelet packet analysis for speaker-independent emotion recognition," *Neurocomputing*, vol. 398, pp. 257–264, Jul. 2020.

SOFIA KANWAL received the M.S. degree in software engineering from International Islamic University, Islamabad, Pakistan, in 2008. She is currently pursuing the Ph.D. degree with COMSATS University Islamabad, under the supervision of Dr. S. Asghar. From 2008 to 2016, she was a Lecturer with the Department of Computer Science, University of Poonch Rawalakot, Azad Kashmir, where she has been promoted to an Assistant Professor, since 2016.

SOHAIL ASGHAR (Member, IEEE) received the degree (Hons.) in computer science from the University of Wales, U.K., in 1994, and the Ph.D. degree from the Faculty of Information Technology, Monash University, Melbourne, Australia, in 2006. In 2011, he joined as the Director with the University Institute of Information Technology, PMAS-Arid Agriculture University, Rawalpindi. He is currently working as a Professor and the Chairman of computer science with COMSATS University Islamabad. He has published extensively (more than 150 publications) in international journals as well as conference proceedings. His research interests include data mining (including structural learning, classification, and privacy preservation in data mining, text, and web mining), big data analytics, data science, and information technology areas. He has also consulted widely on information technology matters, especially in the framework of data mining and data science. He is a member of the Australian Computer Society (ACS) and the Higher Education Commission Approved Supervisor. He has served as a program committee member for numerous international conferences and regularly speaks at international conferences, seminars, and workshops. In 2004, he acquired the Australian Postgraduate Award for Industry. He is in the editorial team of well reputed scientific journals.