Hindawi
Security and Communication Networks
Volume 2022, Article ID 8777026, 10 pages
https://doi.org/10.1155/2022/8777026

Research Article

Ensemble Learning by High-Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal

M. M. Venkata Chalapathi,[1] **M. Rudra Kumar**,[2] **Neeraj Sharma**,[1] **and S. Shitharth**[3]

[1]*School of Engineering, Computer Science and Engineering, Sri Satya Sai University of Technology and Medical Sciences, Sehore, Bhopal, India*
[2]*Department of Computer Science and Engineering, G. Pullaiah College of Engineering and Technology, Kurnool, Andhra Pradesh, India*
[3]*Department of Computer Science and Engineering, Kebri Dehar University, Kebri Dehar 001, Ethiopia*

Correspondence should be addressed to S. Shitharth; shitharths@kdu.edu.et

Received 13 January 2022; Revised 2 February 2022; Accepted 7 February 2022; Published 28 February 2022

Academic Editor: Thippa Reddy G

Copyright © 2022 M. M. Venkata Chalapathi et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In the recent past, handling the high dimensionality demonstrated in the auditory features of speech signals has been a primary focus for machine learning (ML-)based emotion recognition. The incorporation of high-dimensional characteristics in training datasets in the learning phase of ML models influences contemporary approaches to emotion prediction with significant false alerting. The curse of the excessive dimensionality of the training corpus is addressed in the majority of contemporary models. Modern models, on the other hand, place a greater emphasis on merging many classifiers, which can only increase emotion recognition accuracy even when the training corpus contains high-dimensional data points. "Ensemble Learning by High-Dimensional Acoustic Features (EL-HDAF)" is an innovative ensemble model that leverages the diversity assessment of feature values spanned over diversified classes to recommend the best features. Furthermore, the proposed technique employs a one-of-a-kind clustering process to limit the impact of high-dimensional feature values. The experimental inquiry evaluates and compares emotion forecasting using spoken audio data to current methods that use machine learning for emotion recognition. Fourfold cross-validation is used for performance analysis with the standard data corpus.

1. Introduction

Emotions have a profound influence on the physical and psychological well-being in humans. How well patients convey their emotions and how well their therapists recognize and respond to them determine improvement in therapeutic settings. [1] Therapists must deal with enormous volumes of data over a lengthy period of time, which is difficult because they must see numerous patients throughout that time. A platform that can give meaningful speech-based emotion identification insights, for example, might be useful in therapy sessions. EmoViz allows users to take voice samples and use a neural network to determine emotional feelings (such as joyful, sad, angry, surprised, or neutral). Emotional information may be obtained through the examination of spoken audio signals without the need of intrusive technology such as facial recognition or internal signal-based physiological sensor data. Users may view how their emotions have evolved over time and how they have grouped audio and texts based on their emotions using the application EmoViz. [2] Emotion is important in everyday interpersonal connections and is seen as a necessary skill for human communication. [2] It facilitates communication by expressing emotions and responding to individuals being communicated with. Many cognitive and affective computing tasks, such as rational decision-making, perception, and learning, benefit from emotional input. As intelligent systems grow more ubiquitous, emotion identification is becoming increasingly crucial. [3].

Computer games, banking, call centers, video monitoring, and psychiatric diagnosis are just a few examples of real-world applications for emotion detection systems. Other practical applications for emotion detection systems include online learning, business applications, clinical investigations, and entertainment [4, 5]. Voice signals incorporate emotions when it comes to the creation of intelligent systems known as "emotion recognition from speech." Because of a host of intrinsic socioeconomic benefits, speech signals are a great source for emotional computing. Because of their inexpensive cost, they are more appealing for speech emotion recognition research than other physiological signals such as electroencephalograms, electrooculograms, and electrocardiograms [6].

Despite modest development, the accuracy of this approach in identifying fear is lower than for other emotions [7, 8]. When Semwal and colleagues [8] integrated fundamental frequency, ZCR (zero-crossing rate), MFCC, and energy, they were able to identify fear with a 77 percent accuracy. Sun et al. [9] revealed that a deep learning neural network model identified bottleneck information with an accuracy of 62.50 percent in detecting fear.

1.1. Motivation. A number of processes are utilized by machine learning techniques to obtain a collection of speech features that may be used to properly categorize emotions. To build appropriate emotion recognition from a speech system, a suitable collection of characteristics must be chosen from which to train the selected learning algorithm. Emotion recognition algorithms mainly rely on features extracted from spoken audio signals [3, 10]; however, identifying an appropriate feature set is challenging [11]. Speech emotion recognition is challenging for a variety of reasons, including an imperfect description of an emotion and the blurring of the boundaries between distinct emotions. Emotion identification from speech is being improved by introducing new aspects, as demonstrated in [12], with an accuracy of 91.75 percent on an acting corpus when employing PLP characteristics. This accuracy is rather low when compared to the 95.20 percent accuracy attained for the synthesis of acoustic characteristics focusing on MFCC and pitch for recognizing speech emotion. Some studies have sought to agglutinate numerous auditory characteristics to increase the accuracy and precision of speech emotion identification [7, 8].

1.2. Problem Statement. "Ensemble learning" refers to the process of combining various learning models with the goal of producing a better learner as a result. Such algorithms are used in a variety of fields, including medical investigations [13] and dialect prediction [14]. Bagging [15] and boosting [16] are two of the most common ensemble approaches. In terms of accuracy, ensembles of core estimation methods have been shown to outperform single hypotheses [17]. Quinlan [18] tested boosting and bagging ensemble models on a variety of datasets and found them to be remarkably effective. Bagging, as the name implies, aims to train several estimators on random subsets of the dataset. If the training samples are drawn with replacement, they are referred to as "bootstrap samples." Ensemble methods were also used to analyse audio data. Schuller et al. [19] investigated ensemble learning methods for recognizing speaker emotion through speech and found an increase in the accuracy of movie content. Morrison et al. [20] combined several classifiers for emotion recognition tasks in call center practices using an unweighted vote method. However, the majority of the contributions indeed are limited to opt the classification decision delivered by the majority of classifiers used in the ensemble of diversified classifiers. The crux of high-dimensional features remains the same. Hence, rather than the ensemble of diversified classifiers, the focus shall be on handling the high dimensionality of the features.

1.3. Organisation of the Manuscript. This paper's structure includes an introduction to the previously stated ensemble learning by high-dimensional acoustic features for emotion recognition from speech audio signals. In Section 2, we look at related work and numerous models for emotion recognition from speech audio signals. Section 3 covers the methods and materials connected to the suggested model. In Section 4, experimental research is conducted, and the proposed model is compared to other modern models. The conclusion of this article is explained in Section 5, followed by references.

2. Related Work

There have been a few studies on support vector machine ensemble learning [21]. Hu et al. [22] used such an ensemble to solve the problem of rotating machinery failure detection. However, studies of this nature are few and far between.

Bhavan et al. [23] used a bagged ensemble approach on the Emo-DB and achieved a prediction accuracy of 92.45 percent. Shegokar and Sircar [24] proposed a CWT with prosodic elements for recognizing emotion in speech audio signals. Using PCA feature transformation and SVM with quadratic kernel as a classification approach, they achieved an overall accuracy of 60.1 percent on the RAVDESS database. The EMD (empirical mode decomposition) method, which uses the reconstructed signal's optimal features, was used to analyse reconstructed speech signals. On the Spanish database, they were able to achieve an average classification accuracy of 91.16 percent using the RNN technique.

As stated in the introduction, there are numerous reasons why emotion identification remains a major challenge. There is a disconnect between acoustic qualities and human emotions, as well as a theoretical framework for linking voice characteristics to a speaker's emotional state [10, 24–26]. Because of these underlying difficulties, there is

disagreement in the research about which elements are better for recognizing emotion recognition. [10, 26]. When several different types of auditory characteristics are combined, researchers have shown promising results in speech emotion identification [10, 26–29]. They have, however, struggled to find a way to combine the various elements in a way that is both effective and efficient. The study [3, 10, 27] emphasises the importance of identifying appropriate features in order to improve the stability of speech emotion recognition systems. Researchers frequently use specialist software to simplify the extraction, selection, and unification of speech features across multiple sources. Diverse learning algorithms for speaker emotion recognition have been demonstrated to be learned and verified using specific features extracted from public databases.

Multiple neural networks are fused together to achieve the goal of increasing the recognition efficiency from multiple perspectives. When a trained model is applied to an unprepared platform, gradient disappearance and overfitting can easily occur. The ability to generalise is crucial in speech emotion recognition. Ensemble learning has a number of advantages, including the ability to generalise and parallelism. The accuracy and reliability of each individual expert are crucial in ensemble learning [30, 31].

The use of ensemble learning and traditional machine learning approaches in speech emotion recognition has recently increased [32]. Weighted sum fusion was used by Mao et al. [33] to demonstrate that separating complex language features from emotional aspects in speech improves the recognition rate. Liu et al. used a variety of emotional feature subsets to train subclassifiers, which were then used to create a decision-making layer fusion, resulting in improved recognition results. Existing ensemble learning relies heavily on expert credibility allocation, which is a significant flaw in the system. In contrast, the data root for the initial decision is speech features, and acquisition methods are limited, resulting in slight variations across samples and inaccurate grouping information [34, 35].

On this basis, ensemble learning models can be used to make more stable decisions by combining multiple models. On the other hand, each expert's credibility is updated online based on their accuracy rate. Both generalisation and recognition of speech emotions have improved [36].

The most recent attempt to conduct ensemble learning by fusing together diverse categorization strategies was HAF [37], which combined various classification algorithms. Despite the model's superior performance, the high dimensionality of the training corpus remains a problem. This contribution depicts an ensemble learning model for clustering the speech audio signals of the dataset used as input to the training phase to mitigate the negative impact of the high-dimensional features. The suggested method uses the distribution diversity of feature values spanned over different records of divergent emotions to determine the best aspects. The proposed model is motivated by the previously described model titled "Speech Emotion Recognition Using Supervised Bayes Learning on Digital Features (SBL-DF)" [38]. The SBL-DF, on the other hand, does not address the negative impact of high-dimensional features.

3. Methods and Materials

This section explores the materials and methods used in the proposed model that enables to predict emotions in speech audio signals. The materials and methods explored in this section are centric to handle the adverse impact of high-dimensional features towards emotion prediction, feature extraction, feature optimization, and ensemble classification using the adaptive boosting technique as represented in Figure 1. The detailed description of these materials and methods is explored in following sections.

3.1. Dimensionality Reduction. The Fuzzy C-Means [39] clustering technique has been employed to handle the high dimensionality of the training corpus that leads to low sensitivity and specificity, which causes intolerable false-alarming.

The FC-Means method divides the input data $\{r_i \exists r_i \in tC \wedge 1 \le i \le |tC|\}$ into clusters, with each cluster retaining a group of records with a substantial association. Concerning this:

Take the records randomly as centroids and perform fuzzy clustering using Fuzzy C-Means, such that one or more records would be in more than one cluster.

Find the optimal centroids of the resultant clusters and perform the clustering of records recurrently till there is no change in the centroids.

The records that may settle in more than one cluster can be scaled for their relationship by measuring their distance from the centroids of the corresponding clusters having those records.

The algorithm works by distributing membership to each record, resulting in each cluster centroid being proportional to the related format of the distance between each record and the corresponding centroid. The closer the data is to the cluster's centroid, the closer their membership is to a specific core of the cluster. Following the membership iteration, the cluster's centroid shall be revised using the following formulas:

$$\underset{j=1}{\overset{|C|}{\forall}} \left\{ \mu_{ij} = \left[\sum_{k=1}^{|tC|} \left(\frac{\left| c_j \cap r_{ik} \right|}{\left| c_j \right|} \right)^{(2/fi-1)} \right]^{-1} \right\}, \tag{1}$$

$$\underset{j=1}{\overset{|C|}{\forall}} \left\{ c_j = \left\{ \frac{\left(\sum_{i=1}^{|tC|} \left(\mu_{ij} \right)^{fi} * \left| r_i \right| \right)}{\left(\sum_{i=1}^{|tC|} \left(\mu_{ij} \right)^{fi} \right)} \right\} \right\}. \tag{2}$$

The number of records representing the record is indicated by the notation |TC|. The notation c_j reflects the record having aspect with the highest support towards the j^{th} cluster, while the notation $fi \in [1, \infty]$ reveals index fuzziness. Centroids are indicated by the set $C = \{c_1, c_2, \ldots c_{|C|}\}$. The notation μij denotes the Euclidian distance of the i^{th} record of the record $\{r_i \exists r_i \in tC \wedge 1 \le i \le |tC|\}$ towards the current centroid c_j of the j^{th} cluster. The depiction represents the Euclidean distance between the j^{th} cluster centroid and the records of

Figure 1: Data flow diagram of the model.

record $\{r_i \exists r_i \in tC \wedge 1 \le i \le |tC|\}$. This Fuzzy C-Means main algorithm's purpose is fading:

$$J(U,V) = \sum_{i=1}^{|tC|}\sum_{j=1}^{|C|}\left\{(\mu i j)^{fi}\left\|\frac{|c_j \cap r_i|}{|c_j|}\right\|^2 \exists i \le |tC| \wedge j \le |C|\right\}. \quad (3)$$

$|c_j \cap r_{ik}|/|c_j|$ // is the Euclidean distance of the j^{th} cluster centroid c_j as well as the i^{th} record $\{r_i \exists r_i \in tC \wedge 1 \le i \le |tC|\}$.

The steps involved in Fuzzy C-Means clustering are as follows:

(i) The set $tC = \{r_1, r_2, \ldots, r_i, r_{i+1}, \ldots r_{|tC|-1}, r_{|tC|}\}$ represents a set of records such that each record $\{r_i \exists r_i \in tC \wedge 1 \le i \le |tC|\}$ represents the record, whereas the notation $C = \{c_1, c_2, \ldots c_{|C|}\}$ indicates set of centroids of all clusters.

(1) The cluster centroid c_j of the j^{th} cluster has been selected randomly.

(2) The fuzzy membership $\mu i j$ has been computed by utilizing

$$\mu_{ij} = \frac{1}{\sum_{m=1}^{|tC|}\left(|c_j \cap r_{im}|/|c_j|\right)^{(2/fi-1)}}. \quad (4)$$

(3) Here, the fuzzy centroid v_j has been measured by utilizing

$$c_j = \left\{\frac{\left(\sum_{i=1}^{|tC|}(\mu_{ij})^{fi} * |r_i|\right)}{\left(\sum_{i=1}^{|tC|}(\mu_{ij})^{fi}\right)}\right\}. \quad (5)$$

(4) The afore two steps (2&3) are recurrent till the condition $\beta > \|U(m+1) - U(m)\|$ is true or the value of the notation j is minimal.

The notation m in this case reflects the iteration's progress. Criterion termination is indicated by the use of the notation β that ranges between 0 and 1. The notation $U = |C| * (\mu_{ij}) * |tC|$ illustrates a fuzzy membership matrix. Finally, the depiction J denotes the fitness estimation process.

Let the number of fuzzy clusters that have been generated be of the size $|fC|$ of the set $fC = \{fc_1, fc_2, \ldots, fc_{|fC|}\}$, which contains fuzzy clusters in the chronological order.

3.2. Optimal Feature Selection. For each set D_j of the records representing j^{th} the label, find the optimal features (x-coordinates of the given speech audio signal) compared to the counterpart set $\{D_k \exists k \ne j\}$. For each set D_j, a feature (x-coordinate) x_i is optimal if the values projected to the i^{th} set's feature D_j are having distribution diversity with the values projected for the same feature x_i in other sets $\{D_k \exists k \ne j\}$. For each feature x_i of the set D_j, the process shall estimate the diversity weight towards each of the other sets $\{D_k \exists j \ne k\}$, which is the absolute difference between the maximum similarity one and the probable similarity observed $(0 \le p - \text{value} \le 1)$. The mathematical model of identifying optimal features from each pair of sets is portrayed in the following description:

$\forall_{i=1}^{|X|}\{x_i \exists x_i \in X\}$ Begin // for all the feature attributes

$\forall_{j=1}^{(n-1)}\{[x_i^j] \exists [x_i^j] \in D_j\}$// Begin

$dw_{x_i \Rightarrow D_j} = 1$// the overall diversity of the feature x_i concerning the set D_j (label)

$\forall_{k=1}^{(n)}\{[x_i^k] \exists [x_i^k] \in D_k, j \ne k\}$// Begin

$p_{ks} = \text{KS} - \text{test}([x_i^j], [x_i^k])$ // performing the fusion of diversity estimation of the feature x_i between the sets D_j, D_k

$d(x_i)_{D_j \Leftrightarrow D_k} = d\tau$// the diversity $d(x_i)_{D_j \Leftrightarrow D_k}$ of the feature x_i between sets $D_j \Leftrightarrow D_k$ has initialized to distance threshold $d\tau$

if $(p_{ks} < p\tau)$ Begin// the probable similarity value (p_{ks}) observed for the feature x_i between the sets D_j, D_k has found to be greater than the given probability threshold $p\tau$

$d(x_i)_{D_j \Leftrightarrow D_k} = p_{ks}$// the diversity $d(x_i)_{D_j \Leftrightarrow D_k}$ of the feature x_i between sets $D_j \Leftrightarrow D_k$ has been discovered from the ks-test

End //of the condition

$dw_{x_i \Rightarrow D_j} = dw_{x_i \Rightarrow D_j} \otimes d(x_i)_{D_j \Leftrightarrow D_k}$

End // of the iterations

$if\,(dw_{x_i \Rightarrow D_j} \ge d\tau)$ Begin // if the diversity weight $dw_{x_i \Rightarrow D_j}$ of the feature x_i towards the set D_j (label) is greater than or equal to the given diversity threshold $d\tau$,

$fD_j \leftarrow x_i$//then consider the feature x_i is optimal to the set D_j and move that to the optimal features set fD_j

End

End// of the iterations

End // of iterations

// Preprocess the datasets of diversified labels//

$\forall_{j=1}^{lc}\forall_{i=1}^{|X|}\{x_i \exists x_i \in X \wedge x_i \notin fD_j\}$ Begin // for each feature x_i that is selected as an optimal feature of the set D_j of the label j,

$\{D_j\}[x_i]$// discarding the feature x_i and values projected to the corresponding feature from the set D_j

End

3.3. The Classifier

3.3.1. Classification Procedure. This section describes the classifier employed in this proposal, as well as the training stage model and the classification procedure's objective function.

The proposed classifier was built using adaptive boosting. The classifier was designed to combine a large number of Boolean classifiers, also known are weak classifiers, that have been built using decision trees. Each weak classifier was built using the best features taken from a series of quantitative steps. These weak classifiers categorize the provided test data based on whether the condition is true or false. Another bad classifier might label the negatives as bipartite, which includes both false positives and false negatives. This procedure has been repeated until the overall weak classifier determines that the task has been finished. Furthermore, the outcomes obtained, all weak classifiers, in general, are combined into the rating scale and provide the final result.

In this article's projected model, each weak classifier was employed to highlight the coherently ideal features gathered during the quantitative seed phase towards binary classification. The classification technique was also repeated for each risk management implementation using a weak classifier; the corpus component that could not be accurately identified was the focus of the next classifier iteration, known as "boosting." Furthermore, weight classification revealed an inferior classifier, which is employed on each iteration. Completing weak classifiers iteratively results in accurately categorised records from all of these weak classifiers. Each weak classifier, according to the projected method, recommends a certain n-gram for classification accuracy. Furthermore, the classification results of weak classifiers would be justified in order to discover the polarity of the given record. When compared to other binary classification challenges, the Adaboost classifier has been demonstrated to be a feasible approach for optimising DT output (decision trees). It has the potential to be widely employed to improve the performance of various machine learning approaches. The label prediction approach for an unlabelled record consists of the steps listed below:

(i) Extract the values of all considered features from the unlabelled records

(ii) The adaptive boosting classification strategy recommended in this proposal shall be used to predict the germination quality of seed samples as:

(iii) Discover the standard measures of the fitness coefficients of the features towards all weak classifiers

(iv) Consider the values of the features in the given input record; the considered features are optimal in regard to one or more weak classifiers

(v) Prepare the normal distribution for each optimal feature that uses the input value of the feature as a standard measure

(vi) Find the fitness confidences of the input record towards all optimal features of the corresponding weak classifier

(vii) Compare the standard measures of the fitness coefficients discovered during the training phase and fitness confidences of the respective features to predict the label

(viii) There shall be a label assigned to each input record after completing this prediction phase

4. Experimental Study

This section focuses on the proposed model's practical implementation in comparison to some of the latest methods discussed in the literature. This section describes the dataset in detail, the changed programme's requirements, and the system conditions that are critical for performance study. Python [40] is used to execute the model, while PyCharm [41] is used to write the code.

4.1. The Data. For the experimental analysis of the proposed model, the dataset RAVDESS [42] was used, which is a corpus of speech audio signals reflecting a variety of emotions. 247 people who were typical of untrained adult researchers assessed the emotional relevance, expressiveness, and authenticity of the RAVDESS dataset. A total of 72 volunteers have also been made available for the dataset's cross-validation. It has been reported that emotional relevance, reliability, and cross-reliability are all higher. 6204 speech audio signal records were used in the experiment, each of which was labelled with the emotions identified in the corresponding speech audio signal. The following are the counts of records representing different emotions: anger, disgust, fear, joy, neutral, surprise, and sad, where the records labelled as anger, fear, joy, and sad each counted at 1128, disgust counted at 576, neutral counted at 564, and surprise counted at 552. Overall, the 200 words spoken by 200 different people in 200 different emotional contexts represent a wide range of emotions.

4.2. Data Processing. The speech audio signals of the dataset are transmuted into the digital format [43] such that each speech signal transformed to a set of y-coordinates representing the corresponding x-coordinates. It is viewed as a two-dimensional matrix of digital representations of each speech audio signal. A total of seven datasets in the CSV format, each representing one of the emotion labels, are generated after data processing.

4.3. Performance Analysis. This approach has been evaluated for performance using metrics from the confusion matrices of all other contemporary models, including those that use "hybrid acoustic features (HAF)" [37] and "Supervised Bayes Learning of Digital Features (SBL-DF)" [38]. It has to divide the records of each label into two sets to perform cross-validation. The suggested EL-HDAF and contemporary models HAF and SBL-DF have undergone fourfold cross-validation to demonstrate the superiority of EL-HDAF over the existing HAF and SBL-DF models. Table 1

The overall number of records taken for the experimental study is 6204. The records used for training are 4653, and the overall records used for testing are 1551.

In order to evaluate the multilabel cross-validation adopted in the performance analysis, the metrics including precision (positive predictive value) and sensitivity should be used. Some other metrics for analysis that are not deemed significant include the weighted sensitivity,

Table 1: The mean and deviation of the assessment metric values depicted from multifold cross-validation.

Average of 10-fold result and deviations			
Metrics	EL-HDAF	HAF	SBL-DF
Precision	0.944679 ± 0.032751	0.894171 ± 0.058389	0.880646 ± 0.065648
Sensitivity recall	0.954865 ± 0.012566	0.908286 ± 0.010922	0.896128 ± 0.017381
Specificity	0.954897 ± 0.005458	0.907739 ± 0.002584	0.893572 ± 0.010829
F-score	0.951408 ± 0.018341	0.899709 ± 0.031556	0.886927 ± 0.034786
Decision accuracy	0.948429 ± 0.030393	0.894171 ± 0.058389	0.880646 ± 0.065648

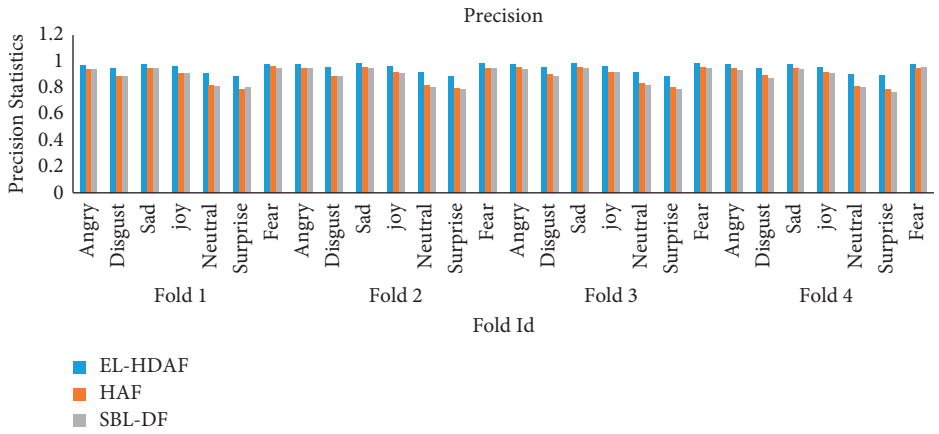

Figure 2: Fourfold cross-validation determined the positive prediction rate (precision).

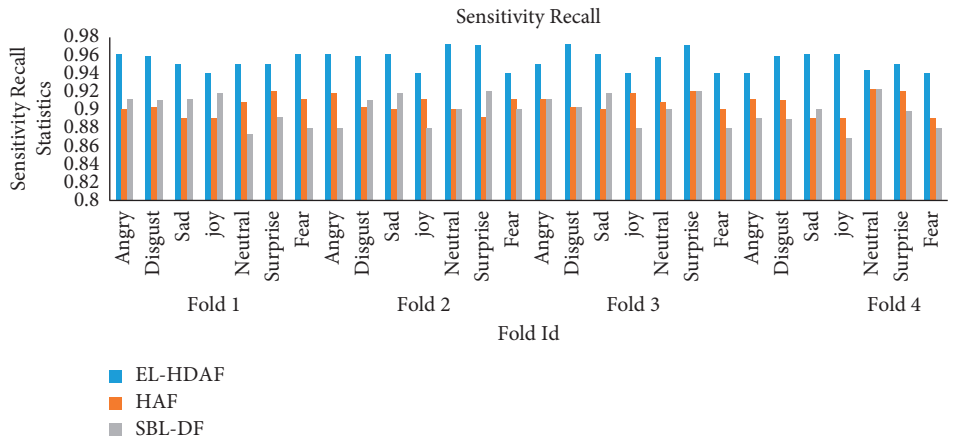

Figure 3: Prediction of emotions sensitivity (recall) as determined by cross-validation, fourfold.

weighted measures of F-score, and precision metrics. The breadth of the solution's implementation and its effectiveness can be determined at the micro-level study of the associated assessment metrics.

When compared to the HAF and SBL-DF approaches, the recommended EL-HDAF strategy shows a more consistent rate of accuracy for all emotions, according to the statistical data shown in Figure 2.

Figure 3 shows that the EL-HDAF has similar performance advantages of emotion prediction sensitivity (recall) compared to contemporaneous models HAF and SBL-DF.

The F-measure and distinct labels are used to display the graphs in Figure 4, with the F-measure representing the harmonic-mean of the precision and sensitivity. The EL-HDAF surpassed the other frameworks, HAF and SBL-DF, that were used for comparison, according to the statistical statistics as in graphical representation.

Figure 5 specifies some factors of which one of the critical measures, the ratio defined for true negative amongst the cumulative set of actual negatives, is considered. The graphical representation of the performance refers to the conditions that refer to the fact that the proposed model is EL-HDAF and is performing superior in comparison to other key models HAF and SBL-DF reviewed for the corpus of requirement specifications. The comparison of the two models is shown in the form of graph with the help of fourfold labels as angry, disgust, fear, glad, neutral, sad, and surprised. Thus, it has been concluded that the performance of the proposed model in terms of specificity is better in all the labels while compared to the contemporary models.

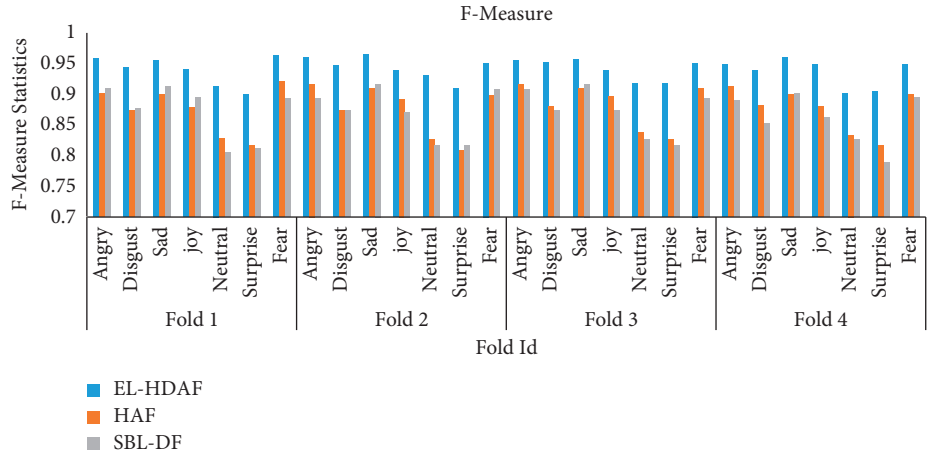

Figure 4: Fourfold cross-validation of EL-HDAF, HAF, and SBL-DF contributed a mean.

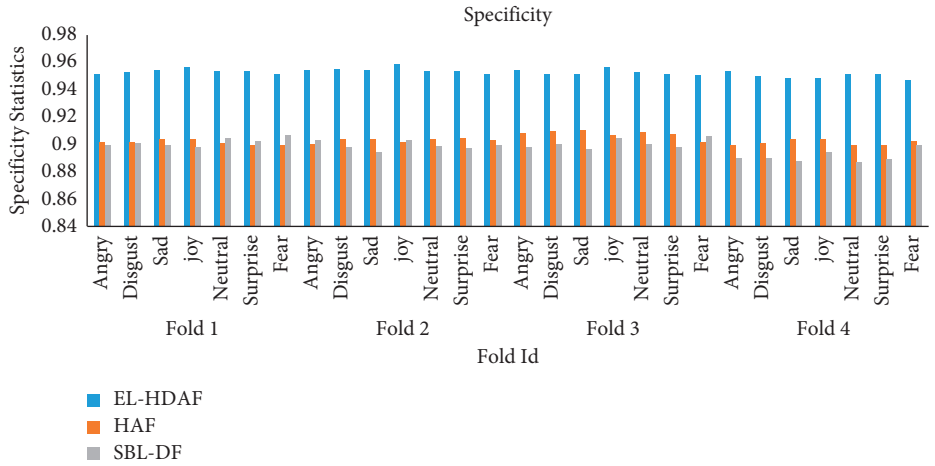

Figure 5: Specificity observed for the proposed EL-HDAF and contemporary models HAF and SBL-DF in terms of metric specificity over fourfolds.

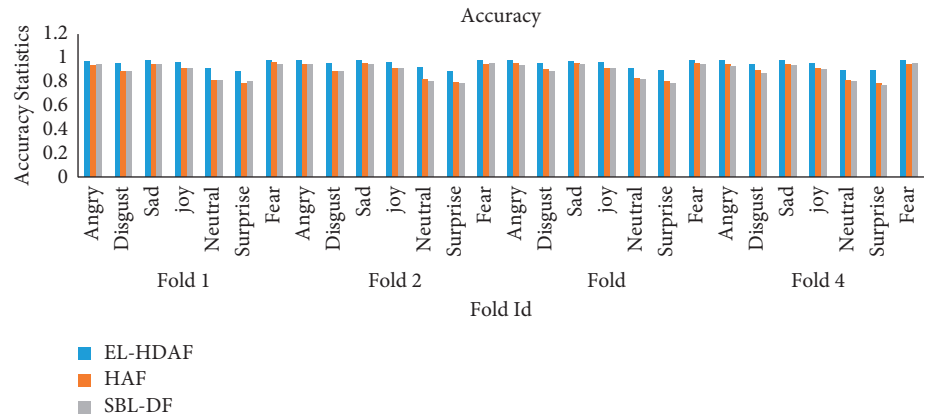

Figure 6: Accuracy of EL-HDAF, HAF, and SBL-DF over fourfolds.

The accuracy metric has been used for measuring the performance of EL-HDAF, HAF, and SBL-DF over the fourfolds as exhibited in Figure 6. The comparison of the three models is shown in the form of graph with the help of fourfold labels as angry, disgust, fear, glad, neutral, sad and surprised. Therefore, it has been concluded that the performance of the proposed model in terms of accuracy is better in all the labels compared to other contemporary models.

Weighted measures of accuracy, recall, and F-score are all essential metrics in determining the strength of the

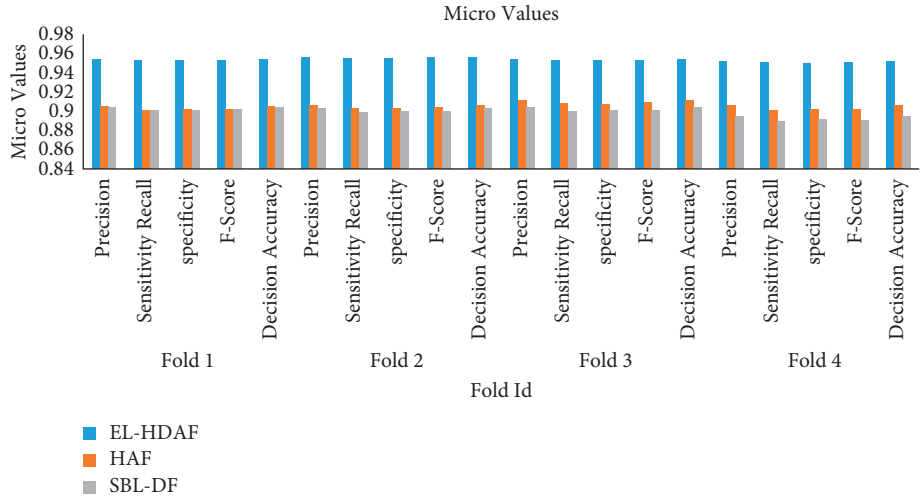

Figure 7: Micromeasures of precision, sensitivity (recall), f-measure, as well as accuracy.

multilabel classification performance because they assist to understand the classifier's performance overall. The metric values represent the classifier's ability to scale its performance based on the precision, sensitivity, and evaluation accuracy factors that include the harmonic-mean of the precision and sensitivity. The micromeasures of the corresponding metrics precision, sensitivity, accuracy, and f-score are also critical to assess the performance of multilabel classification.

For EL-HDAF, HAF, and SBL-DF, the weighted measures of the corresponding metrics observed for each emotion prediction are the essential inputs to determine the micromeasures of the corresponding metrics. The micromeasures of the corresponding cross-validation metrics are represented in Figure 7. The fourfold cross-validation process and the resultant micromeasures of precision, sensitivity, f-score, and class prediction accuracy indicate that the model EL-HDAF outperforms the models SBL-DF and HAF.

5. Conclusion

In recent years, predicting emotional states from acoustic features of spoken audio signals has been a prominent objective in the field of speech audio signal processing. Machine learning models with a high feature dimension are used to recognize empathy from audio data. To reduce the effect of high-dimensional data on the proposed model during training, the feature values of various classes were analysed for diversity, and a novel clustering approach was devised. It is also worth mentioning that the adaptive boosting classification technique is intended to learn from the various clusters in the training corpus. Ensemble Learning by High-Dimensional Acoustic Features (EL-HDAF) is a projected model that has been evaluated against two existing models, HAF and SBL-DF, using the benchmark dataset RAVDESS using fourfold cross-validation. In performance analysis, the cross-validation metrics and accompanying micromeasures were investigated. The results of the suggested and current measurements demonstrate that EL-emotion HDAF detection beats the existing methods HAF and SBL-DF with the fewest false alarms and the highest decision accuracy. In the future, the acoustic features of the speech stream can be adjusted utilizing evolutionary computing methodologies to increase the performance of ensemble learning models in predicting emotion. The contribution would motivate future research towards emotion detection through acoustic features of speech signals, where an evolutionary technique has an optimal scope in feature optimization.

Abbreviations

ML: Machine learning
EL-HDAF: Ensemble learning by high-dimensional acoustic features
ZCR: Zero-crossing rate
EMD: Empirical mode decomposition
HAF: Hybrid acoustic features
SBL-DF: Speech emotion recognition using supervised Bayes learning on digital features
c_j: Cluster centroid
μij: Euclidean distance
v_j: Fuzzy centroid
$|fC|$: Fuzzy clusters
D_j: Diversity
x_i: Feature
(p_{ks}): Probable similarity value
$p\tau$: Probability threshold
$dw_{x_i \Rightarrow D_j}$: Diversity weight
DT: Decision trees.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request (shitharths@kdu.edu.et).

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Authors' Contributions

M M Venkata Chalapathi conceptualised the study, curated the data, performed a formal analysis, devised the methodology, contributed to the software, and wrote the original draft; M. Rudra Kumar supervised the study, wrote and reviewed the content, edited the article, and helped with project administration and visualization; Neeraj Sharma supervised the study, wrote and reviewed the software, validated the content, and wrote the original draft and was responsible for devising the methodology; S. Shitharth wrote, reviewed, and edited the article, helped acquire funding, and contributed to the visualization and formal analysis, and also software development.

References

[1] D. Fabiano and S. Canavan, "Emotion recognition using fused physiological signals," in *Proceedings of the 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 42–48, IEEE, Cambridge, United Kingdom, September 2019.

[2] Y. Ma, Y. Hao, M. Chen, J. Chen, P. Lu, and A. Košir, "Audio-visual emotion fusion (AVEF): a deep efficient weighted approach," *Information Fusion*, vol. 46, pp. 184–192, 2019.

[3] K. Sarker and K. R. Alam, "Emotion recognition from human speech: emphasizing on relevant feature selection and majority voting technique," in *Proceedings of the 3rd International Conference on Informatics, Electronics & Vision (ICIEV)*, pp. 89–95, Dhaka, Bangladesh, May 2014.

[4] R. Subhashini and P. R. Niveditha, "Analyzing and detecting employee's emotion for amelioration of organizations," *Procedia Computer Science*, vol. 48, pp. 530–536, 2015.

[5] A. Rychalski and S. Hudson, "Asymmetric effects of customer emotions on satisfaction and loyalty in a utilitarian service context," *Journal of Business Research*, vol. 71, pp. 84–91, 2017.

[6] M. Papakostas, E. Spyrou, T. Giannakopoulos et al., "Deep visual attributes vs. hand-crafted audio features on multi-domain speech emotion recognition," *Computation*, vol. 5, no. 2, p. 26, 2017.

[7] A. Khan and U. K. Roy, "Emotion recognition using prosodie and spectral features of speech and Naïve Bayes Classifier," in *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 1017–1021, IEEE, Chennai, India, March 2017.

[8] N. Semwal, A. Kumar, and S. Narayanan, "Automatic speech emotion detection system using multi-domain acoustic feature selection and classification models," in *Proceedings of the 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pp. 1–6, IEEE, New Delhi, India, February 2017.

[9] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on DNN-decision tree SVM model," *Speech Communication*, vol. 115, pp. 29–37, 2019.

[10] I. Luengo, E. Navas, and I. Hernaez, "Feature analysis and evaluation for automatic emotion identification in speech," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 490–501, 2010.

[11] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 109–114, IEEE, Coimbatore, India, March 2017.

[12] H. K. Palo, M. Chandra, and M. N. Mohanty, "Emotion recognition using MLP and GMM for Oriya language," *International Journal of Computational Vision and Robotics*, vol. 7, no. 4, pp. 426–442, 2017.

[13] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," *Computer Methods and Programs in Biomedicine*, vol. 104, no. 3, pp. 443–451, 2011.

[14] X. Sun, "Pitch accent prediction using ensemble machine learning," in *Proceedings of the Seventh International Conference on Spoken Language Processing*, Denver, CL, USA, September 2002.

[15] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[16] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, 1990.

[17] H. J. V. Veen, *Le Nguyen the Dat Armando Segnini*, Kaggle Ensembling Guide, 2015.

[18] J. R. Quinlan, "Bagging, boosting, and C4. 5," in *Proceedings of the the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, pp. 725–730, Portland, Oregon, August 1996.

[19] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker independent speech emotion recognition by ensemble classification," in *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo*, pp. 864–867, IEEE, Amsterdam, Netherlands, July 2005.

[20] D. Morrison, R. Wang, and L. C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[21] H. C. Kim, S. Pang, H. M. Je, D. Kim, and S. Y. Bang, "Support vector machine ensemble with bagging," *Pattern Recognition with Support Vector Machines*, Springer, in *Proceedings of the International Workshop on Pattern Recognition with Support Vector Machines*, pp. 397–408, August 2002.

[22] Q. Hu, Z. He, Z. Zhang, and Y. Zi, "Fault diagnosis of rotating machinery based on improved wavelet package transform and SVMs ensemble," *Mechanical Systems and Signal Processing*, vol. 21, no. 2, pp. 688–705, 2007.

[23] A. Bhavan, P. Chauhan, R. R. Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," *Knowledge-Based Systems*, vol. 184, Article ID 104886, 2019.

[24] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *Proceedings of the 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–8, IEEE, Surfers Paradise, Gold Coast, Australia, December 2016.

[25] S. Parthasarathy and I. Tashev, "Convolutional neural network techniques for speech emotion recognition," in *Proceedings of the2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 121–125, IEEE, Hitotsubashi Hall in Tokyo, Japan, September 2018.

[26] W. Jiang, Z. Wang, J. S. Jin, X. Han, and C. Li, "Speech emotion recognition with heterogeneous feature unification of deep neural network," *Sensors*, vol. 19, no. 12, p. 2730, 2019.

[27] Z. T. Liu, M. Wu, W. H. Cao, J. W. Mao, J.-P. Xu, and G. Z. Tan, "Speech emotion recognition based on feature

selection and extreme learning machine decision tree," *Neurocomputing*, vol. 273, pp. 271–280, 2018.
[28] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, 2015.
[29] J. B. Alonso, J. Cabrera, M. Medina, and C. M. Travieso, "New approach in quantification of emotional intensity from the speech signal: emotional temperature," *Expert Systems with Applications*, vol. 42, no. 24, pp. 9554–9564, 2015.
[30] W. Zehra, A. R. Javed, Z. Jalil, H. U. Khan, and T. R. Gadekallu, "Cross corpus multi-lingual speech emotion recognition using ensemble learning," *Complex & Intelligent Systems*, vol. 7, pp. 1–10, 2021.
[31] O. Obulesu, K. Suresh, D. Gaurav et al., "Adaptive diagnosis of lung cancer by deep learning classification using wilcoxon gain and generator," *Journal of Healthcare Engineering*, vol. 2021, Article ID 5912051, 13 pages, 2021.
[32] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, pp. 47–58, 2021.
[33] Q. Mao, X. Zhao, and Y. Zhan, "Extraction and analysis for non-personalized emotion features of speech," *Advances in Information Sciences and Service Sciences*, vol. 3, no. 10, pp. 225–263, 2011.
[34] S. Shitharth, B. M. Gouse, R. Kadiyala, and B. Vidhyacharan, *Prediction of COVID-19 Wide Spread in India Using Time Series Forecasting Techniques*, Springer, Berlin, Germany, 2021.
[35] G. T. Reddy, S. Bhattacharya, S. S. Ramakrishnan et al., "An ensemble based machine learning model for diabetic retinopathy classification," in *Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE)*, pp. 1–6, IEEE, Vellore Institute of Technology, Vellore, India, February 2020.
[36] A. K. Cherian, E. Poovammal, N. S. Philip, K. Ramana, S. Singh, and I. H. Ra, "Deep learning based filtering algorithm for noise removal in underwater images," *Water*, vol. 13, no. 19, p. 2742, 2021.
[37] K. Zvarevashe and O. Olugbara, "Ensemble learning of hybrid acoustic features for speech emotion recognition," *Algorithms*, vol. 13, no. 3, p. 70, 2020.
[38] M. V. Chalapathi, "Speech emotion recognition using supervised Bayes learning on digital features of multi-label data corpus," *Design Engineering*, pp. 1065–1078, 2021.
[39] M. Shasidhar, V. S. Raja, and B. V. Kumar, "MRI brain image segmentation using modified fuzzy c-means clustering algorithm," in *Proceedings of the 2011 International Conference on Communication Systems and Network Technologies*, pp. 473–478, IEEE, Katra, India, June, 2011.
[40] "Python. (n.d.),".
[41] "pycharm. (n.d.),".
[42] S. R. 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, Article ID e0196391, 2018.
[43] "wav-to-csv. (n.d.),".