ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR872

Enhancements in Immediate Speech Emotion Detection: Harnessing Prosodic and Spectral Characteristics

ZEWAR Shah [1]; SHAN Zhiyong [*1]; Adnan [2]
[1] College of Information Science and Technology, Donghua University, Shanghai 201620, P.R China
[2] School of Computer Science and Technology Donghua University Shanghai 201620, China

Abstract:- Speech is essential to human communication for expressing and understanding feelings. Emotional speech processing has challenges with expert data sampling, dataset organization, and computational complexity in large-scale analysis. This study aims to reduce data redundancy and high dimensionality by introducing a new speech emotion recognition system. The system employs Diffusion Map to reduce dimensionality and includes Decision Trees and K-Nearest Neighbors（KNN）ensemble classifiers. These strategies are suggested to increase voice emotion recognition accuracy. Speech emotion recognition is gaining popularity in affective computing for usage in medical, industry, and academics. This project aims to provide an efficient and robust real-time emotion identification framework. In order to identify emotions using supervised machine learning models, this work makes use of paralinguistic factors such as intensity, pitch, and MFCC. In order to classify data, experimental analysis integrates prosodic and spectral information utilizing methods like Random Forest, Multilayer Perceptron, SVM, KNN, and Gaussian Naïve Bayes. Fast training times make these machine learning models excellent for real-time applications. SVM and MLP have the highest accuracy at 70.86% and 79.52%, respectively. Comparisons to benchmarks show significant improvements over earlier models.

Keywords:- *Feature Extraction, KNN, Speech Emotions, Diffusion Map, MFCC, and Feature Engineering.*

I. INTRODUCTION

Speech is one of the most basic and pervasive forms of human communication, speech is capable of expressing and understanding a broad range of emotions. Emotional speech processing is an interdisciplinary area of study that tries to understand and analyze the subtle emotional undertones in spoken language [1]. This task calls for a wide range of computational approaches and resources, all with the common goal of extracting and analyzing emotional indicators embedded in voice signals. Emotional voice processing has numerous potentials uses in many fields, such as virtual assistants, speech recognition, customer service, mental health diagnostics, education, and human-robot interactions. Researchers have devoted a lot of time and energy over the last 20 years to trying to decipher the complex signs that speech uses to convey emotion [2]. In response to this need, several models based on machine learning have emerged with the express purpose of extracting and categorizing emotional content from audio recordings. At its core, our effort revolves around feature extraction, the first and most important stage in classifying emotional states in spoken words. The significance of feature engineering in classification is shown by the fact that the effectiveness of speech identification systems is dependent on the quantity and quality of these derived features [3].One of the main obstacles to automated emotion recognition in speech is the complex nature of acoustic feature extraction. Regardless, numerous studies have tried to overcome these obstacles by developing speech emotion recognition systems that use machine learning paradigms, all with the ultimate aim of better comprehending human emotions [4]. The intricate nature of human communication makes it extremely difficult to incorporate speech emotion detection into larger computer frameworks. In addition to being able to recognize and understand human emotions in speech, intelligent computer systems should be able to mimic the subtle behavioral reactions seen in real-life interactions. Consequently, there are a plethora of machine learning techniques that may be employed to tackle the complex problem of voice emotion detection categorization. The overarching goal of this research is to advance our understanding of human emotional expressions in computational linguistics by delving into the many tactics and methodologies used to build strong speech emotion detection mechanisms [5]. Precise and computationally efficient classification is essential for the effectiveness of speech emotion recognition (SER) systems. Developing features involves using strict selection criteria, estimating feature number, and implementing dependable extraction processes, which are crucial for this endeavor [6]. Developers of SER models should carefully consider the development procedure of the features as it impacts the temporal complexity and classification accuracy. Identifying reliable emotional cues in datasets necessitates having access to extensive and diverse repositories that encompass a broad spectrum of emotions and languages [7]. The main objectives of emotion recognition (SER) [8] are to (1) define the semantic basis of emotions and (2) determine the speech signal components or parameters that are important for emotional expression. Research in this subject has highlighted the significance of feature aspects such quality metrics, prosodic qualities, spectral features, and voice-related variables in determining the efficiency of SER systems [9-11]. Developing effective SER frameworks requires more than just employing robust feature selection; it also entails utilizing sophisticated classification methods [12]. Machine learning techniques for feature engineering and categorization have gained increased attention from the academic community. The three most commonly utilized classifiers in SER [13] are K-Nearest Neighbor (KNN), Decision Trees (DT), and Support Vector Machine (SVM), each offering distinct advantages and considerations during model development [14]. Rosalind Picard's pioneering research in emotive computing had a significant impact on Speech Emotion Recognition (SER), which is crucial for HMI. The utilization of physical cues, such as facial expressions, speech, and body position, enhances the field of human-computer interaction. Two potential applications include the detection of driver drowsiness and the dissemination of public education. The study proposes a system that integrates ensemble classifiers with dimension reduction, specifically emphasizing paralinguistic attributes such as intensity and pitch for speech and emotion recognition (SER). The emotion categorization process involves the utilization of the Ryerson Audio-Visual Database of Emotional Speech and Song, in conjunction with machine learning techniques such as Support Vector Machine and Gaussian Naïve Bayes. This study emphasizes the need of carefully choosing distinguishing characteristics and enhancing the performance of models by adjusting hyper parameters. It underscores the crucial role played by feature selection and classification methodologies. This study aims to enhance the accuracy and efficiency of emotion identification in speech utterances by addressing the issues of redundancy and high dimensionality associated with human feature extraction. This study accomplishes its objectives by implementing a non-linear approach for feature extraction using a principled technique based on learning models that utilize Diffusion Maps. The proposed method, utilizing Mel Frequency Cepstral Coefficients (MFCC) for feature extraction together with advanced techniques such as batch normalization and z-score standardization, achieves a 3.1% higher classification accuracy compared to conventional genetic algorithms. The rest of this academic paper is structured as follows: Section 2 provides an extensive literature analysis, Section 3 outlines the methodology and data sources utilized, Section 4 displays the experimental results and subsequent discussion, and Section 5 closes by synthesizing the results and proposing future study directions.

II. LITERATURE REVIEW

When it comes to Speech Emotion Recognition (SER), optimization solutions are crucial, especially when it comes to feature extraction and selection. The key goal is to find the best solutions as quickly as possible without sacrificing accuracy. Specifically, Yoon et al. sought to determine the most important characteristics in speech data for emotion detection and proposed a particle swarm optimization method for feature selection in SER. On the other hand, another study used deep convolutional neural networks to extract spectrogram properties; nevertheless, their performance was limited because they solely optimized solutions without taking additional contextual aspects into account.

Contrarily, Kanwal et al. suggested using DBSCAN and PCA to improve the precision of speech emotion recognition, specifically for the diverse types of emotions described in the INTERSPEECH 2010 feature set. Their strategy outperformed current state-of-the-art approaches in terms of accuracy and recall. On the other hand, problems surfaced throughout testing, such as using the wrong feature sets and using inefficient feature engineering techniques, which increased computational costs and resulted in less-than-ideal convergence rates.

Previous studies have attempted to address the time complexity of SER optimization; for example, [15] aimed to decrease processing time by using a smaller feature set and a one-class-in-one neural network, a type of specialized neural network. Optimal emotion prediction using a subset of Linear Predictive Coding (LPC) and Delta LPC parameters was the basis of this strategy, which sought to simplify computational operations. Still, it has been observed that these approaches' reliance on indirect comparisons increases processing time and requires big datasets for validating results effectively [16]. To conclude, SER optimization solutions provide great potential for improving classification accuracy and computational efficiency. However, to guarantee dependable performance in real-world applications, it is crucial to address inherent limitations like biases in feature selection and indirect comparisons. A careful approach to feature comparison is needed to avoid high-dimensional speech emotion feature datasets that are prone to feature redundancy. This is because of the complex relationship between individual characteristics in speech data. Feature data with too many dimensions causes training periods to be longer and puts a load on computational resources. A key to overcoming these obstacles is zeroing down on the most important aspects of speech that convey different emotions. Feature dimensionality reduction techniques have been the subject of several investigations [17], but large-scale access to professional labeled datasets is required for the development of effective classification systems.

The authors have come up with a new method of extraction and reduction to address these issues; this method has potential in many other areas, including pattern recognition and speech analysis. The Diffusion Map (DM) strategy, which uses Mel Frequency Cepstral Coefficients (MFCC) in conjunction with dimensionality reduction and exhaustive feature selection methodologies, provides a new way to improve the extraction of speech emotion data. Applying the benchmark EMO-DB dataset, we test DM's effectiveness across seven different emotion classes and find encouraging numerical results that point to its supremacy. Some argue that a large variety of features is necessary to capture the complex range of emotions expressed in speech data, while others argue that a smaller set of features is more appropriate for this purpose. Feature extraction and selection are promising ways to get around data dimensionality problems, and adding Independent Data projection techniques opens up even more possibilities for improving classification accuracy. Evaluation of ML models for speaker-independent emotion recognition from voice data highlights the importance of strong feature engineering techniques for pushing the boundaries of speech emotion recognition. The effectiveness of a system in the field of bimodal information Speech Emotion Recognition (SER) is heavily dependent on the features that are input. Criteria controlling feature extraction, whether using model-driven or data-driven learning approaches, are inextricably bound up with the categorization process. To overcome these difficulties caused by different language settings, a new method has been suggested that uses cross-corpus multilingual Ensemble learning supported by a majority voting strategy [3]. Applying this strategy to three separate ML algorithms—Sequential Minimal Optimization (SMO), Random Forest (RF), and J48—proved to significantly boost WA and UA on the Urdu, EMO-DB, and SAVEE (Surrey Audio-Visual Expressed Emotion) datasets.

The availability of corpora covering several languages in naturalistic contexts is still low, despite progress [2]. In an effort to decipher emotional signals in audio signals and frequency parameterizations, researchers have investigated various feature extraction methodologies, such as Mel-Gammatone-Frequency Cepstral Coefficients (MGFCC), Gammatone-Frequency Cepstral Coefficients (GFCC), and Mel Frequency Cepstral Coefficients (MFCC). Although there are some limitations, such as the dataset's categorical representation of emotional classes and difficulties in speech interpretation and data annotation, the results of analyses performed on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset were promising [4]. Innovative spectral feature modulation methods have been suggested to overcome these restrictions [18, 7]. These methods include Random Forest (RF) classifiers, Support Vector Machine (SVM) combinations with Single Frequency Filtering (SFF), and Linear Discriminant scoring systems. Utilizing SFF to extract the amplitude envelope of the speech signal in conjunction with classifiers like SVM and RF has significantly improved emotion detection rates on multiple datasets, such as EMO-DB, FAU-AIBO, and IEMOCAP [4][2]. But there are still obstacles, especially when it comes to identifying happy feelings because there aren't enough language and auditory clues [4]. The novel framework Bidirectional Long-Short-Term Memory with Directional Self-Attention (BLSTM-DSA) addresses these concerns. Incorporating Bi-Directional mechanisms into encoding and decoding processes may improve emotion identification systems in bimodal environments and diversify information representations. Multi-class SVM, Random Forest (RF), and Single-layered Adaptive Boosting were applied to SAVEE and Polish datasets to improve Speech Emotion Recognition (SER). Random forest models using smaller feature sizes reduced processing time and attained testing accuracies of 75.71% and 87.91% on these datasets [11]. Using 3D Mel Spectrograms, Parallel AG-TFNN, and 2D Attention Gated Tensor Neural Network architecture, we explored further SER design developments. These architectures—augmented with LSTM+CNN baselines using Adam optimizer—performed differently across emotion classes on the EMO-DB and IEMOCAP datasets in speaker-independent modes [7]. Both datasets showed promising results, but independent testing highlighted issues with the 2D and 3D representations' dimensionality and redundancy, requiring further parameter adjustments [3]. To address emotion threshold extraction's high dimensionality and redundancy, Diffusion Map-based learning algorithms were proposed. Non-linear feature extraction allowed this strategy to outperform random selection genetic algorithms. A diffusion map-based non-linear technique boosted accuracy by 3.1%, achieving an impressive 89.01% accuracy rate when extracting experimental features from the EMO-DB dataset that are independent of speakers [3]. This unique approach revolutionizes SER methodology by laying the groundwork for extracting complex emotional elements and eliminating redundancy and high dimensionality. This approach could advance Speech Emotion Recognition in affective computing, HCI, and clinical diagnostics with rigorous testing and improvement. Using a highly deep CNN to create extra feature maps for training acoustic models, the author in [23] discovered that the enriched data aids in the development of strong voice recognition systems. Using a blend of Gaussian distributions as the random source, the author synthesized feature vectors through an AAE in [24]. Although synthetic samples have the potential to enhance classification performance, they often adhere to a random distribution instead of the actual data distribution. Their approach for creating synthetic feature vectors was also constructed using convolutional neural networks (cGANs). The cGAN has been trained using a number of training methods, including initializing the generator with the weights from the AAE decoder [25] and updating the generator's weights multiple times before updating the discriminator in each training epoch. Finding an optimal trade-off between the generator's and discriminator's capabilities is a big challenge while training GANs. Applying dynamic alternation training can help you overcome this challenge. In this training method, instead of fixing the number of training epochs between the discriminator and generator, it is dynamically changed. Our suggested network's goal is to enhance learning stability by making it easier for the generator to learn the target distribution, rather than optimizing the number of training epochs. Our suggested network's generator, in particular, doesn't learn a fixed distribution but rather the distribution of a latent representation generated by a concurrently learned encoder.

III. MATERIAL & METHOD

This section provides an extensive synopsis of the suggested approach, covering the architecture of feature extraction with a DM for fitness assessment included. In addition, the dataset, classification algorithms, and DM's application are discussed in depth, with Figure 1 serving to clarify the former.

A. Extraction of Feature s

The proposed framework integrates characteristics acquired from several sources, encompassing intensity, pitch, LFCCs, MFCCs, formants, and spectral centroids. Non-linear pitch perception is captured by MFCCs through the representation of acoustic data on the Mel scale. The comprehension of high-frequency vocal tract stimulation can be enhanced by the utilization of low-frequency coulombic filters (LFCCs), which exhibit variations in the linear scale of their band-pass filters. Resonances in the vocal tract are represented by formants, while spectra centroids display the weighted average of sound frequencies. Pitch serves as a representation of the fundamental frequency, whereas intensity denotes the magnitude of sound strength per unit area. In order to account for the dynamics of speech, characteristics of velocity and acceleration are included for MFCCs, LFCCs, intensity, and pitch. The feature vector with a dimensionality of 652 is obtained by multiplying the feature extraction process using the Librosa and Parselmouth packages. These programs employ statistical methods to combine feature vectors from various signal durations.

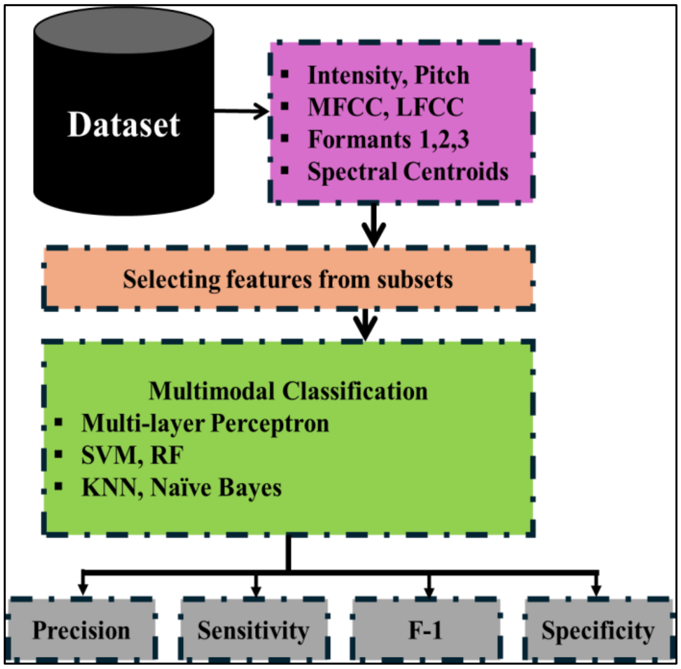


Fig 1. Developed Methodology

B. Selecting Features from Subsets

The architecture of machine learning models strongly depends on the selection of feature subsets to enhance classifier performance and reduce dimensionality. The Filter, Wrapper, and Embedded approaches are among the most often employed techniques. Wrapper methods involve the iterative evaluation of subsets, which can be computationally expensive. In contrast, filter approaches are independent of the classifier. Embedded strategies integrate feature selection into the model itself to address the limitations of both filter and wrapper methods. Several techniques employed in recent research on speech emotion recognition encompass swarm-based optimization, Boruta, principal component analysis (PCA), forward and backward selection, among others. This work utilizes a filter-based technique to enhance statistical dependency on important features by leveraging mutual knowledge, owing to its non-model-specific nature, simplicity, and efficiency.

The methodology reduces the initial collection of 1582 characteristics to a more manageable set of 90 features, which are crucial for classification, by computing a family of embeddings in Euclidean space drawn from the eigenvectors and eigenvalues of a diffusion operator. While research into the specific features that promote efficient data clustering into different categories is continuing, these extracted features are crucial for training models to determine the output labels that are sought. An opportunity for additional optimization and refinement within the framework of the suggested methodology is to test a large number of different features and incorporate them into a standardized feature set. To fix α € R and a rotation invariant kernel.

$$k_\varepsilon(x,y) = h(\frac{||x-y||2}{\varepsilon})$$

Now Suppose that;

$$S\varepsilon_{(x)} = \int K\varepsilon(x,y)s(y)dy$$

New kernel is

$$k_\varepsilon^\alpha(x,y) = K\varepsilon(x,y)/S(x)S(y)$$

Apply the weighted graph Laplacian normalization to this kernel by setting.

$$d\varepsilon_{(x)} = \int aK\varepsilon(x,y)s(x)dy$$

Anisotropic transition kernel:

$$t_\varepsilon^x(x,y) = k_\varepsilon^\alpha(x,y)/td_\varepsilon^x(x)$$

C. Classification Method (Naïve Bayes)

The penultimate stage is to use five machine learning models to categorize speech into eight different moods, after feature extraction and subset selection. To evaluate the models' accuracy in both overall and class-wise terms, they are evaluated individually using Stratified k-fold cross-validation (k = 10). For its execution, the study makes use of the scikitlearn package. Assuming conditional independence among features, the Gaussian Naïve Bayes Classifier, one of the models, depends on the Bayes Theorem. Although it is computationally simple, the assumption of independence leads to suboptimal performance because it is seen unreasonable. Stratified tenfold cross-validation with all 652 features yielded a classifier accuracy of 43.48 percent. The accuracy rate was 44.98% when just 177 features were considered.

Table 1. Selection of feature and features subset

Features Description	*Total Features*	*Selected Features*
Intensity	18	18
Pitch	18	18
LFCCs	234	47
Spectral Centroids	4	4
MFCCs	360	72

D. Classification Method (KNN)

A simple machine learning method, k-Nearest Neighbors (k-NN) stores all training instances without instantly learning from them, using a lazy-learner approach. To find the k locations that are geographically closest to the test case, the classification algorithm uses the Euclidean distance. A grid search determines the ideal value of k, which can be anything from 1 to 15, according to the inquiry. By setting k to 1, stratified tenfold cross-validation was able to attain an overall accuracy of 53.25%. The classifier distributes weight equally among all neighbors. The overall accuracy is improved to 65.17 percent by using a subset of 177 characteristics, all hyperparameters are kept at their original levels. This enhancement is noteworthy.

E. Classification Method (RF)

Ensemble learning is a technique that combines predictions from numerous lesser-performing models in order to create a resilient learner. Random forests, which consist of ensembles of decision trees, utilize a voting process that relies on majority to determine the final forecasts. Bootstrapping is employed to generate individual trees, while grid search is utilized to determine split criteria such as Gini impurity and entropy. The performance of random forests generally exhibits enhancement as the number of decision trees increases, particularly when considering Gini impurity, hence indicating a higher level of split quality. With 100 trees and Gini impurity, the RF classifier attains an overall accuracy of 61.84%. The classifier's accuracy increases to 64.06% when it is trained on a specific subset of 177 characteristics.

F. Classification Method (SVM)

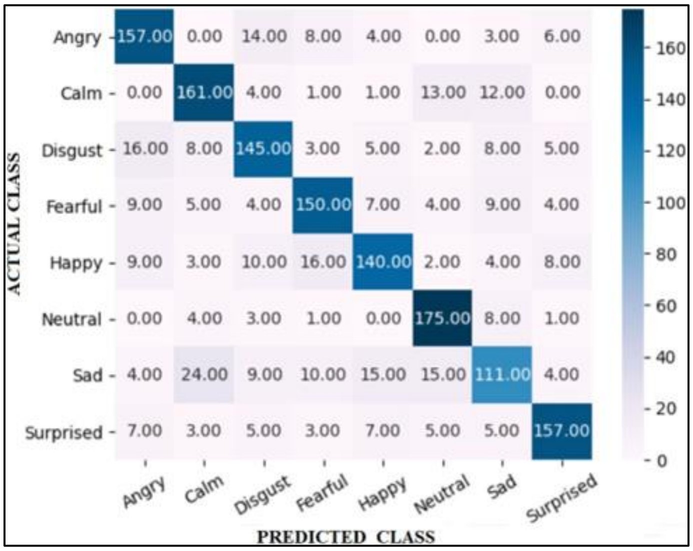
A strong binary classifier, Support Vector Machines (SVMs) maximize margins by finding a hyperplane between two classes. Using the kernel trick, support vector machines (SVMs) build this hyperplane in a space with more dimensions. The training points are made linearly separable by selecting appropriate kernel functions. This research makes use of support vector machines (SVMs) with Gaussian kernels, which are defined by RBF kernels. The optimal values for the punishment term (C) and the kernel coefficient (γ) were determined to be 100 and 0.0029, respectively, by employing stratified tenfold cross-validation and grid search, respectively. Using these hyperparameters, the model achieves an impressive 67.57% accuracy with all 652 features and 70.86% accuracy with the subset of 177 features, demonstrating a significant boost in performance.

G. Classification Method (MLP)

A multilayer perceptron (MLP) is made up of layers of neurons, with each layer being represented by a node. Each layer computes its output using a linear combination of inputs, with the addition of nonlinearity introduced by an activation function (σ). Sigmoid, ReLU, and tanh are some of the most common activation functions. The two-hidden-layer MLP used in this research was trained using the Adam optimizer for twenty iterations with categorical cross-entropy serving as the loss function. When all characteristics are used in a stratified tenfold cross-validation model, the overall accuracy is 73.76%. The performance is improved to 79.62% accuracy by using the subset of 177 features that were chosen.

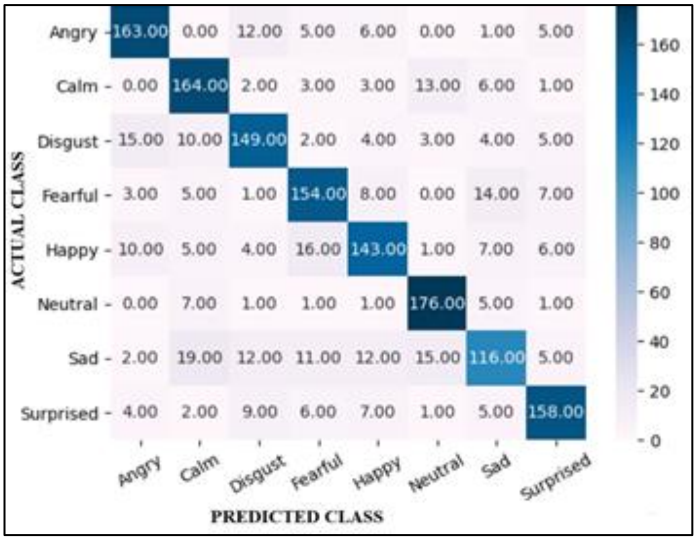
IV. RESULTS

This part encompasses an overview of the implementation procedure, as well as a comprehensive analysis of the study's findings and outcomes. The dataset utilized in this study is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). It comprises a total of 1440 audio files, encompassing the expressions of 24 performers across 8 distinct emotions, each characterized by two intensity levels. The Librosa library in Python is utilized to read the data, followed by the extraction of features using Librosa and Parselmouth. Additionally, as part of the data preparation process, the data is scaled to conform to a Standard Normal distribution. Feature selection, which relies on mutual information, retains the top 20% of LFCC and MFCC features. The Synthetic Minority Oversampling Technique (SMOTE) was employed to address the class imbalance, resulting in a total of 1536 audio files, with 192 files matching each emotion. In the process of training and validating models, stratified k-fold cross-validation is employed to ensure that the subsets utilized for training and evaluation accurately reflect the entirety of emotions. In this study, the classification models are assessed utilizing a stratified tenfold cross-validation methodology.



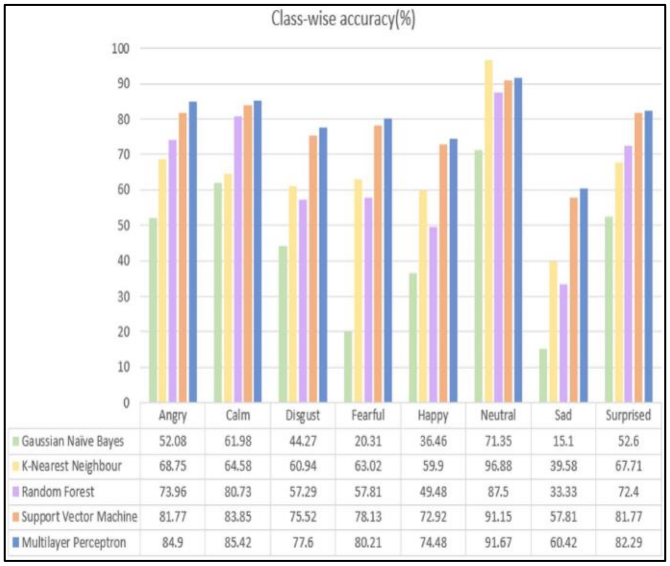
	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprised
Angry	157.00	0.00	14.00	8.00	4.00	0.00	3.00	6.00
Calm	0.00	161.00	4.00	1.00	1.00	13.00	12.00	0.00
Disgust	16.00	8.00	145.00	3.00	5.00	2.00	8.00	5.00
Fearful	9.00	5.00	4.00	150.00	7.00	4.00	9.00	4.00
Happy	9.00	3.00	10.00	16.00	140.00	2.00	4.00	8.00
Neutral	0.00	4.00	3.00	1.00	0.00	175.00	8.00	1.00
Sad	4.00	24.00	9.00	10.00	15.00	15.00	111.00	4.00
Surprised	7.00	3.00	5.00	3.00	7.00	5.00	5.00	157.00

Fig 2. SVM Confusion Matrix



	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprised
Angry	163.00	0.00	12.00	5.00	6.00	0.00	1.00	5.00
Calm	0.00	164.00	2.00	3.00	3.00	13.00	6.00	1.00
Disgust	15.00	10.00	149.00	2.00	4.00	3.00	4.00	5.00
Fearful	3.00	5.00	1.00	154.00	8.00	0.00	14.00	7.00
Happy	10.00	5.00	4.00	16.00	143.00	1.00	7.00	6.00
Neutral	0.00	7.00	1.00	1.00	1.00	176.00	5.00	1.00
Sad	2.00	19.00	12.00	11.00	12.00	15.00	116.00	5.00
Surprised	4.00	2.00	9.00	6.00	7.00	1.00	5.00	158.00

Fig 3. MLP Confusion Matrix



	Angry	Calm	Disgust	Fearful	Happy	Neutral	Sad	Surprised
🟩 Gaussian Naive Bayes	52.08	61.98	44.27	20.31	36.46	71.35	15.1	52.6
🟨 K-Nearest Neighbour	68.75	64.58	60.94	63.02	59.9	96.88	39.58	67.71
🟪 Random Forest	73.96	80.73	57.29	57.81	49.48	87.5	33.33	72.4
🟧 Support Vector Machine	81.77	83.85	75.52	78.13	72.92	91.15	57.81	81.77
🟦 Multilayer Perceptron	84.9	85.42	77.6	80.21	74.48	91.67	60.42	82.29

Fig 4. Different Emotion Classifier Performance

This section provides a summary of the implementation process, along with a thorough examination of the study's results and consequences. This study employs the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) as the dataset. The collection consists of 1440 audio files, which include the vocalizations of 24 artists representing 8 different moods, each with two levels of strength.

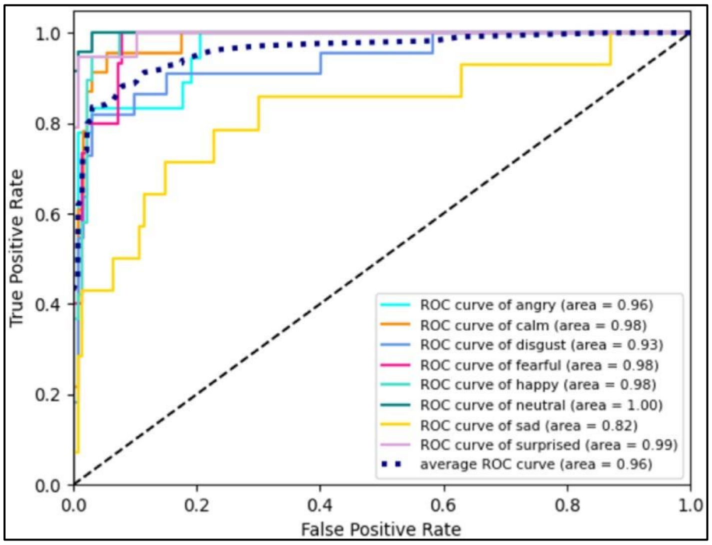


Fig 5. MLP Model ROC Plots

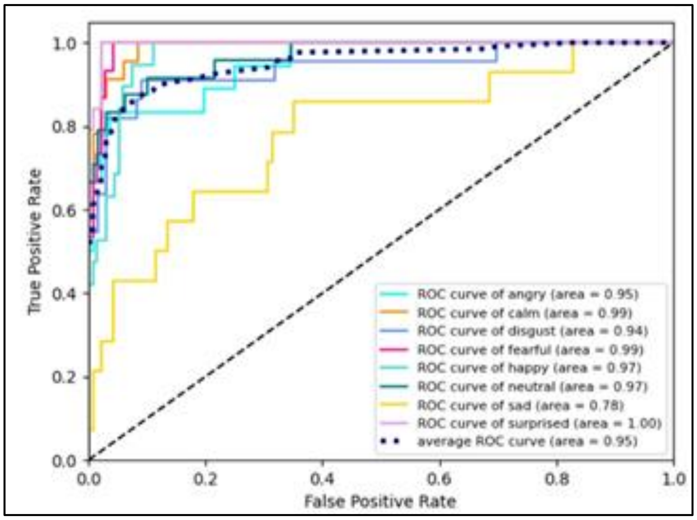


Fig 6. SVM Model ROC Plots.

The data is read using the Librosa library in Python, and subsequently, features are extracted using Librosa and Parselmouth. Furthermore, in the context of data preparation, the data undergoes scaling to ensure its conformity to a Standard Normal distribution. Feature selection selects the top 20% of MFCC and LFCC characteristics based on mutual information. To rectify the class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was utilized, yielding a grand total of 1536 audio files, with 192 files corresponding to each emotion. Stratified k-fold cross-validation is used during the training and validation of models to verify that the subsets used for training and evaluation accurately represent all emotions. This study employs a stratified tenfold cross-validation methodology to evaluate the categorization models.

Table 2. Developed Methodology Comparison with Existing State of the Art.

Ref	Deployed Model with Article reference Number	Accuracy	Dataset
[1]	Features optimization with DGA+PCA [1]	77.49%	EMO-DB, RAVDESS
[4]	SMO, RF, J48, Ensemble Learning [4]	69.31%	URDU (SMOEL), EMODB (SMOEL), SAVEE (RFEL), EMOVO(SMOEL)
[5]	TLFMRF [5]	68.89%	CASIA Corpus, Emo-DB
[7]	T-FNN, 2D and 3D (AG-TFNN), Parallel, AGTFNN [7]	70.09%	Emo-DB, IEMOCAP
[19]	SFF and nonlinear energy operator, SVM, RF [19]	-	EMODB, FAUAIBO, IEMOCAP
[6]	iMEMD, SVM and k-NN [6]	71.70%	SEED, DEAP
[8]	D-MFCC [8]	86.96%	CASIA
[14]	EMD and its second order difference plots (SODP) SVM [14]	-	DEAP
Developed Methodology	DM+DT+KNN [Developed Methodology]	89.01%	EMO-DB

For speaker-independent fitness evaluation, our research presents a new feature extraction architecture that combines a dimension reduction mechanism (DM) with Decision Tree (DT) and K-Nearest Neighbor (KNN) classifiers. This method outperforms methods that use random selection to evaluate fitness, with an emphasis on using a non-linear approach based on diffusion maps for feature extraction. To make emotion recognition more reliable, we use the Mel Frequency Cepstral Coefficients (MFCC) feature selection approach to extract features from audio signals and frequency parameterization. Results from our speaker-independent studies showed an average accuracy of 88.82% with a range of 86.93% to 89.01%; compared to earlier results from the EMO-DB dataset, our suggested methodology showed a significant 3.1% improvement in accuracy. In addition, Table 2 presents extensive findings, including emotion-wise accuracy for speaker-independent scenarios, so you may see how well our suggested approach works across different emotional categories.

V. DISCUSSION

Our SER framework will be compared to other SER frameworks in the literature, including SVM and MLP models, to see how well it performs. Table 2 compares the SVM and MLP models' overall accuracy with that of existing SER models published for the RAVDESS dataset in the literature. Tenfold stratified cross-validation was employed to compile and present the findings of our study. A study in [20] employed support vector machines (SVMs) with different kernel functions to classify data using characteristics obtained from continuous wavelet transform (CWT). The maximum reported accuracy for the RAVDESS dataset was 60.1% author in [21] attained an accuracy of 64.48 percent by utilizing spectrograms in combination with a classifier that relies on deep neural networks. To train a Logistic Model Tree, the authors in [22] use a 13-dimensional feature vector made up of MFCCs. A 70% accuracy rate is their claimed maximum overall performance. Keep in mind that MFCCs can't pick up on every nuance of human emotion in speech. The spectrum features, in particular the delta and delta-delta coefficients, determine the exact classification of emotions. In comparison to the previously mentioned models on the RAVDESS dataset, our suggested models perform better. Because a human observer can only achieve a 67% accuracy rate when using the RAVDESS dataset, it poses a serious challenge to emotion detection researchers [23]. Researchers used a bagged ensemble of SVMs to identify MFCCs and spectral centroids in their work [24]. The researchers achieved an impressive performance of 75.69% on the RAVDESS dataset using this approach and a train-test split ratio of 90:10. As a point of comparison, our best result with MLP is 84.96%, and with SVM, it's 86.27%. To train a convolutional neural network to determine the emotional state of spoken language, researchers in [25] used a mix of MFCCs and modulation spectral characteristics. They achieved a maximum accuracy rate of 78.10%. It is noteworthy that the Support Vector Machine (SVM) model, which achieves an accuracy of 77.86%, is a favorable choice because to its comparatively shorter training time, while being a strong contender among the recommended classification strategies. Consequently, the SER framework employed in the study demonstrates superior performance compared to previously reported frameworks in terms of overall accuracy ratings. The SVM system we propose is computationally efficient and highly dependable, as it keeps just 177 features after selecting a subset of features. The recommended Support Vector Machine (SVM) framework is highly suitable for inclusion into real-time speech emotion recognition applications due to its shorter training periods in comparison to its Multilayer Perceptron (MLP) counterpart.

VI. CONCLUSION

The study's Speech Emotion Recognition model makes use of machine learning algorithms like k-Nearest Neighbours, Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP). Analysis of spectral and prosodic features leads to the selection of 177 features among 652 overall, based on the information they share with the target attribute. Feature subset selection techniques based on filters are utilised in real-time applications because to their computational efficiency. Results demonstrate that MLP and SVM classifiers both achieve better validation accuracy than others, even though SVM offers shorter training durations. Results from statistical tests show that SVM outperforms MLP in real-time scenarios, making it a good substitute for MLP. Adding audio quality elements and adjusting MFCC and LFCC coefficients are potential areas for future study to improve system performance. Further research into gender-based SER systems and alternative feature selection algorithms might lead to useful improvements.

Feature engineering and classification are the main phases in speech emotion recognition (SER). Our research introduces a novel dimensionality and feature reduction method utilising the Ryerson Audio-Visual Database of Emotional Speech and Song dataset, outperforming baseline methods. The SER model, which uses the suggested dimensionality reduction technique, is compared to various state-of-the-art studies for recall and accuracy; Table II shows the results. By incorporating numerous sets of characteristics at the data classification level, the SER model that follows the suggested reduction and extraction strategy performs similarly. Fusion weights help us understand how each feature set affects SER performance and their magnitude. In earlier research without gender consideration, standalone KNN classifiers have not consistently given optimal results. Ensemble approaches, especially AdaBoost, perform better even when they just employ a subset of the training samples or attributes. This shows the features of subsets' value in new ensemble classification methods. Our preliminary Emo-Berlin dataset results suggested that our ensemble approaches and state-of-the-art results may be improved. A binary-based classifier that displays confusion matrices may enhance SER accuracy. We anticipate exploring these alternatives may improve SER approaches and boost emotion recognition performance. In recognizing emotions from speech patterns, ensemble, decision tree (DT), and K-Nearest Neighbor (KNN) classifiers and diffusion map-based machine learning have showed promising results. Diffusion maps can help us understand complex speech feature interactions and improve emotion recognition algorithms. Finally, the simple and effective DT and KNN classifiers make it easy to categorize emotions using extracted voice data. The diffusion map-based technique extracts data structure and reduces feature space dimensionality. The nearest neighbor network integrates easily with pairwise-based classification models and improves computing efficiency, especially with high-dimensional data. Redundancy can be reduced by using features with negligible effect on accuracy, dimensionality reduction, and feature selection. Different people can react differently to the same speech, especially in real-life situations without a script, making emotion identification harder. A robust Speech Emotion Recognition (SER) system with diverse datasets to train models and increase accuracy is needed to handle this complexity. Future research employing the diffusion map-based method and neural-based approaches could increase diagnostic accuracy and efficiency by analysing clinical images and facial expressions to learn about health issues and diseases.

REFERENCES

[1]. S. Kanwal and S. Asghar, “Speech Emotion Recognition using Clustering Based GA-Optimized Feature Set”, IEEE access, vol. 9, pp. 125830-125842, 2021.

[2]. R. Hidayat, “Frequency Domain Analysis of MFCC Feature Extraction in Children’s Speech Recognition System”, JURNAL INFOTEL (Informatics, Telecommunication, and Electronics), vol. 14, no. 1, pp. 30-36, 2022.

[3]. S. Yildirim, Y. Kaya and F. Kılıç, “A Modified Feature Selection Method Based on Metaheuristic Algorithms for Speech Emotion Recognition”, Applied Acoustics, vol. 173, 107721, 2021.

[4]. M. Swain, A. Routray and P. Kabisatpathy, “Databases, Features and Classifiers for Speech Emotion Recognition: A Review”, International Journal of Speech Technology, vol. 21, pp. 93-120, 2018.

[5]. Z. Zhang, “Mechanics of Human Voice Production and Control”, The Journal of the Acoustical Society of America, vol. 140, no. 4, pp. 2614-2635, 2016.

[6]. M. B. Akçay and K. Oğuz, “Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, Supporting Modalities, and Classifiers”, Speech Communication, vol. 116, pp. 56-76, 2020.

[7]. N. Salankar, P. Mishra and L. Garg, “Emotion Recognition from EEG Signals using Empirical Mode Decomposition and Second-Order Difference Plot”, Biomedical Signal Processing and Control, vol. 65, 102389, 2021.

[8]. R. Thirumuru, K. Gurugubelli and A. K. Vuppala, “Novel Feature Representation using Single Frequency Filtering and Nonlinear Energy Operator for Speech Emotion Recognition”, Digital Signal Processing, vol. 120, 103293, 2022.

[9]. C. K. 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 Systems with Applications, vol. 69, pp. 149-158, 2017.

[10]. 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”, Procedia Computer Science, vol. 49, pp. 50-57, 2015.

[11]. C. C. Lee, E. Mower, C. Busso, S. Lee and S. Narayanan, “Emotion Recognition using a Hierarchical Binary Decision Tree Approach”, Speech Communication, vol. 53, no. 9-10, pp. 1162-1171, 2011.
[12]. K. S. Rao, S. G. Koolagudi and R. R. Vempada, “Emotion Recognition from Speech using Global and Local Prosodic Features”, International Journal of Speech Technology, vol. 16, pp. 143-160, 2013.
[13]. S. Prasomphan and S. Doungwichain, “Detecting Human Emotions in a Large Size of Database by using Ensemble Classification Model”, Mobile Networks and Applications, vo.. 23, pp. 1097-1102, 2018.
[14]. M. B. Mustafa, M. A. M. Yusoof, Z. M. Don and M. Malekzedeh, “Speech Emotion Recognition Research: An Analysis of Research Focus”, International Journal of Speech Technology, vol. 21, pp. 137-156, 2018.
[15]. S. Yoon, S. Byun and K. Jung, “Multimodal Speech Emotion Recognition using Audio and Text”, IEEE Spoken Language Technology Workshop (SLT), 18-21 December, 2018, Greece, pp. 112-118.
[16]. E. Bingham and H. Mannila, “Random Projection in Dimensionality Reduction: Applications to Image and Text Data”, Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 26-29 August, 2001, California, pp. 245-250.
[17]. 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, e0196391, 2018.
[18]. S. R. Kadiri and P. Alku, “Excitation Features of Speech for Speaker-Specific Emotion Detection”, IEEE Access, vol. 8, pp. 60382-60391, 2020.
[19]. B. Schuller, R. Muller, M. Lang and G. Rigoll, “Speaker Independent Emotion Recognition by Early Fusion of Acoustic and Linguistic Features Within Ensembles”, 9th European Conference on Speech Communication and Technology, 04-08 September, 2005, Portugal, pp. 1-4
[20]. Shegokar P, Sircar P. Continuous wavelet transform based speech emotion recognition. In2016 10th International conference on signal processing and communication systems (ICSPCS) 2016 Dec 19 (pp. 1-8). IEEE.
[21]. Zeng Y, Mao H, Peng D, Yi Z. Spectrogram based multi-task audio classification. Multimedia Tools and Applications. 2019 Feb;78:3705-22.
[22]. Zamil AA, Hasan S, Baki SM, Adam JM, Zaman I. Emotion detection from speech signals using voting mechanism on classified frames. In2019 international conference on robotics, electrical and signal processing techniques (ICREST) 2019 Jan 10 (pp. 281-285). IEEE.
[23]. de Lope J, Grana M. An ongoing review of speech emotion recognition. Neurocomputing. 2023 Apr 1;528:1-1.
[24]. Ye J, Wen XC, Wei Y, Xu Y, Liu K, Shan H. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition. InICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2023 Jun 4 (pp. 1-5). IEEE.
[25]. Jha T, Kavya R, Christopher J, Arunachalam V. Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. International Journal of Speech Technology. 2022 Sep;25(3):707-25.

AUTHORS

First Author – Zewar Shah, Collage of Information science and Technology, Donghua University, Shanghai 201620, P.R China.
Second Author –Shan Zhiyong, Collage of Information science and Technology, Donghua University, Shanghai 201620, P.R China.
Third Author –Adnan, *School of Computer Science and Technology Donghua University* Shanghai 201620, China

Correspondence Author – **Shan zhiyong,** School of Information and Technology, Donghua University, Shanghai 201620, P.R China