African Journal of Advanced Pure and Applied Sciences (AJAPAS)

Online ISSN: 2957-644X

Volume 3, Issue 3, 2024, Page No: 454-468

Website: *https://aaasjournals.com/index.php/ajapas/index*

معامل التأثير العربي 2023: (1.55)   SJIFactor 2023: 5.689   ISI 2022-2023: 0.557

Special issue: First Libyan Conference on Technology and Innovation (LCTI-2024), Benghazi, Libya

Impact of Dimension Reduction Techniques on the Accuracy of Speech Emotion Recognition

Reem, M, Ben-Sauod [1]*, Rayhan, S, Alshwihde [2], Wafa, I, Eltarhouni [3]

[1,2,3] Department of Computer Science, Faculty of Information Technology, University of Benghazi, Benghazi, Libya

تأثير تقنيات تقليل الأبعاد على دقة التعرف على عواطف الكلام

ريم محمد بن سعود[1]*، ريحان سليمان الشويهدي[2]، وفاء الترهوني[3]

[3،2،1] علوم الحاسوب، كلية تقنية المعلومات، جامعة بنغازي، بنغازي، ليبيا

*Corresponding author: reem.bensauod@uob.edu.ly

Received: March 01, 2024	Accepted: May 05, 2024	Published: May 30, 2024

Abstract:

Dimensionality reduction techniques play an important role in the accuracy of speech emotion recognition (SER). This research focuses on the utilization of feature selection (FS) and feature reduction (FR) techniques for SER. The proposed approach introduces a ConvLSTM model that combines convolutional neural networks (CNN) and long short-term memory (LSTM) networks to improve the accuracy of SER. . The study investigates the effect of the four Dimensional Reduction (DR) techniques chosen for this experiment. Extracted four acoustic features Mel-frequency cepstral coefficients (MFCC), Mel-spectrogram, Chromagram, Root Mean Square (RMS). Data augmentation (DA) techniques are applied to enhance the model's robustness and introduce variations to the training data. The effectiveness of FS and FR techniques is evaluated using three widely employed audio datasets. The experimental results demonstrate the superiority of the proposed approach over previous studies. The correlation-based feature selection (CFS) technique achieved the highest accuracy rates of 97.22% for RAVDESS, 96.65% for SAVEE, and 97.75% for EMO-DB. Similarly, in the 4-fold cross-validation, CFS achieved high accuracy rates ranging from 97.11% to 98.39%. Additionally, in FR techniques was the Principal Component Analysis (PCA) technique performed well for feature reduction. The results underscore the importance of selecting appropriate feature selection and reduction techniques based on factors such as dataset type, size, and compatibility with subsequent models.

Keywords: SER, feature selection, feature reduction, LSTM, CNN, DA**.**

الملخص

تلعب تقنيات تقليل الأبعاد دورًا مهمًا في دقة التعرف على عواطف الكلام (SER). ركز هذه الورقة البحثية على استخدام تقنيات اختيار الميزات (FS) وتقليل الميزات (FR) للتعرف على المشاعر من إشارة الكلام يقدم النهج المقترح نموذج ConvLSTM الذي يجمع بين الشبكات العصبية التلافيفية (CNN) وشبكات الذاكرة طويلة المدى (LSTM) لتحسين دقة SER. تبحث الدراسة في تأثير تقنيات تخفيض الأبعاد الأربعة (DR)المختارة لهذه التجربة. تم استخراج أربع ميزات صوتية: معاملات ميل التردد الرأسي (MFCC)، مخطط ميل الطيفي، مخطط كروماجرام، مربع متوسط الجذر (RMS)، يتم تطبيق تقنيات زيادة البيانات (DA) لتعزيز قوة النموذج وإدخال الاختلافات في بيانات التدريب. يتم تقييم فعالية تقنيات FS وFR باستخدام ثلاث مجموعات بيانات صوتية متاحة للجمهور ومستخدمة على نطاق واسع. وتبين النتائج التجريبية تفوق المنهج المقترح على الدراسات السابقة. حققت تقنية اختيار الميزات المستندة إلى الارتباط (CFS) أعلى معدلات دقة بلغت 97.22% لـ RAVDESS، و96.65% لـ SAVEE، و97.75% لـ EMO-DB. وبالمثل، في التحقق المتبادل بأربعة أضعاف، حققت CFS معدلات دقة عالية تتراوح من 97.11% إلى 98.39%. بالإضافة إلى ذلك، في تقنيات FR، كانت تقنية PCA تؤدي أداءً جيدًا لتقليل الميزات. تؤكد النتائج على أهمية اختيار التقنيات المناسبة لاختيار الميزات وتقليلها بناءً على عوامل مثل نوع مجموعة البيانات وحجمها والتوافق مع النماذج اللاحقة.

الكلمات المفتاحية: SER، اختيار الميزات، تقليل الميزات، LSTM، CNN، DA.

Introduction

SER has gained significant attention in affective computing due to its crucial role in various applications, including human-computer interaction, speech-based therapy, and sentiment analysis [1]. Accurate detection and classification of emotions from speech signals can improve the quality of human-machine interactions and enable personalized and adaptive systems [2]. However, speech data often contains a large number of features, resulting in computational complexity. Additionally, not all features contribute equally to the emotion recognition task, and irrelevant or redundant features may introduce noise and degrade classification model performance. The challenges posed by high dimensionality are noteworthy. Firstly, individuals encounter difficulty in conceptualizing and interpreting high-dimensional spaces, making it challenging to gain intuitive insights from models. Secondly, algorithms face obstacles in learning meaningful patterns when the number of input data sources surpasses the available training data [3]. To address these challenges, dimensionality reduction and feature selection techniques are employed to reduce noise and extract relevant information from complex inputs. These techniques play a vital role in dimensionality reduction, contributing to the improvement of accuracy and interpretability in speech emotion recognition systems [4].

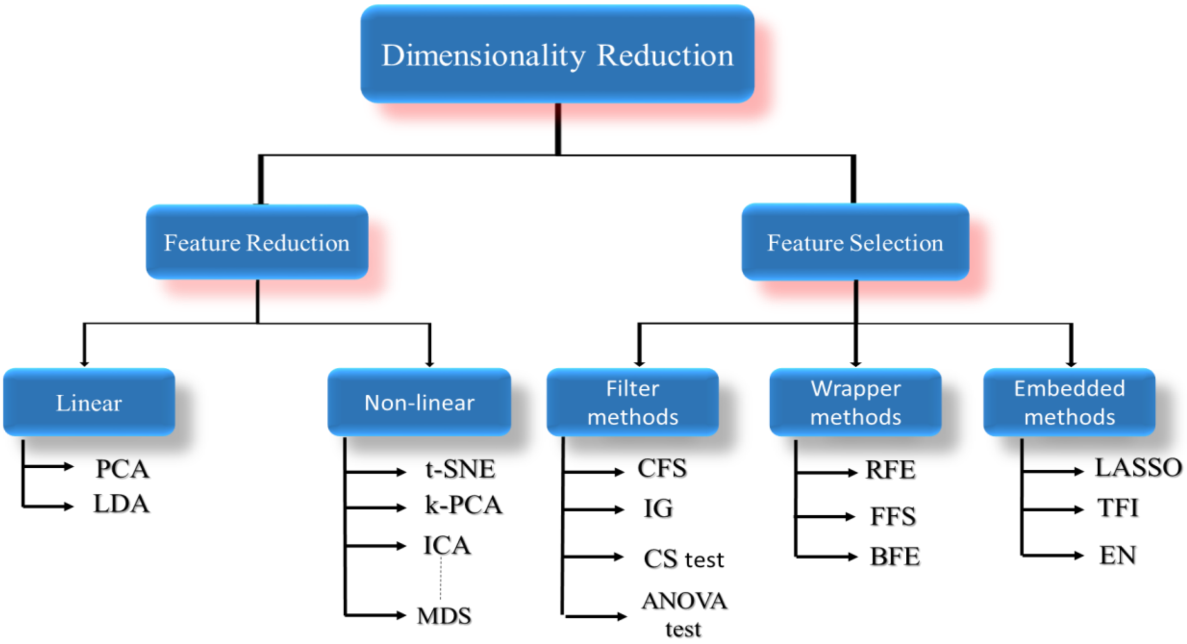


Figure 1: Dimension reduction techniques.

DR as shown in Figure 1 aim to transform high-dimensional feature spaces into lower-dimensional representations while preserving essential information. By mitigating the challenges associated with dimensionality, these techniques enhance the generalization capabilities of classification models and reduce computational complexity [5].

FR, it is a technique commonly used in machine learning and data mining to minimise data complexity and increase model performance. PCA, Linear Discriminant Analysis (LDA), and t-distributed Stochastic Neighbour Embedding (t-SNE) are commonly utilized methods in speech emotion recognition. Linear methods, such as PCA and LDA, utilize linear transformations to extract significant features from high-dimensional data. PCA identifies principal components that account for the maximum variance within the data, while LDA focuses on finding projections that maximize the separability between different classes [6] [7].

FS, also known as variable selection or attribute selection, are techniques used to identify and select a subset of relevant features from a larger set of available features or variables. The goal of feature selection is to improve the performance of machine learning models by reducing the dimensionality of the input space and focusing on the most informative and discriminative features. Feature selection methods can be categorized into three main types: filter methods, wrapper methods, and embedded methods [8].

- Filter methods evaluate the relevance of features independently of any specific learning algorithm. These methods assess the individual characteristics of features, such as their correlation with the target variable or their statistical properties. Popular filter methods include: CFS, information gain (IG), Chi-square test (CS test), mutual information, Analysis of variance test (ANOVA test) [9].

- Wrapper methods select features by incorporating a specific learning algorithm. They assess subsets of features by training and evaluating the model's performance using different feature combinations. Wrapper methods tend to be computationally more expensive compared to filter methods. Examples of wrapper methods include: recursive feature Elimination (RFE), forward feature selection (FFS), backward feature elimination (BFE) [10].
- Embedded methods perform feature selection as part of the model training process. These methods select features based on their importance derived from the internal mechanisms of the learning algorithm. Embedded methods are often specific to certain algorithms and include techniques such as: L1 regularization (LASSO), Tree-based feature importance (TFI), elastic net (EN) [10].

This research aims to investigate and evaluate the impact of FS and FR techniques on the accuracy of speech emotion recognition. By applying different FS algorithms and FR methods on diverse speech datasets, we aim to identify the most effective techniques that enhance the classification performance and improve the efficiency of emotion recognition systems. The findings of this research can provide valuable insights into the selection and pre-processing of speech features for emotion recognition and contribute to the development of more robust and efficient affective computing systems.

In the following sections, the related work in the field will be discussed, the methodology employed in this research will be presented, the experimental setup will be described, and the results obtained from the application of various feature selection and dimension reduction techniques on speech emotion recognition datasets will be analysed in the following sections. Finally, a summary of the findings, their implications, and potential directions for future research in this domain will be provided.

Related Work

The effectiveness of FS and FR methods in enhancing the efficiency of speech emotion recognition systems has remained uncertain. To address this gap, this literature review focuses on discussing and evaluating the utilization of feature selection techniques for data augmentation in the context of speech emotion recognition systems.

The research paper [11], presents an improved system for SER through the extraction of a hybrid high-dimensional feature vector from speech and glottal-waveform signals. The feature extraction stage incorporates various techniques, including MFCC, Perceptual Linear Predictive Coding (P LPC), Minimum Variance Distortionless Response (MVDR), and prosodic features derived from the fundamental frequency contour. To achieve DR and estimate Gaussian Mixture Model (GMM) classifier parameters, the proposed system employs a modified quantum-behaved particle swarm optimization (QPSO) algorithm known as pQPSO. The pQPSO algorithm is specifically designed to efficiently search within a limited parameter range. The system's performance is evaluated on three widely recognized emotional speech databases. Results indicate that the proposed system exhibits improved accuracy compared to several classical methods commonly employed in dimension reduction, such as Factor Analysis (FA), PCA, PPCA, and LDA. Specifically, the proposed system achieves accuracies of 82.82%, 60.79%, and 74.80% for the EMO-DB, SAVEE, and IEMOCAP datasets, respectively. In comparison, classical dimension reduction techniques yield accuracies of 67.36% for LDA, 66.16% for PCA 83.33% for FA, and 77.08% for PPCA. In this study [9], investigates the utilization of a deep convolutional neural network (DCNN) to leverage its advantages in speech emotion recognition SER. Specifically, a pre-trained framework is applied to extract features from speech emotion databases. Additionally, a feature selection FS approach is implemented to identify the most discriminative and important features for SER. The classification task is performed using various algorithms, including random forest (RF), decision tree (DT), support vector machine (SVM), multilayer perceptron classifier (MLP), and k-nearest neighbors (KNN). The experiments are conducted on four publicly accessible databases. The proposed method demonstrates high accuracies in recognition when employing the feature selection technique CFS. Accuracies of 92.02%, 88.77%, 93.61%, and 77.23 are achieved for Emo-DB, SAVEE, RAVDESS, and IEMOCAP, respectively. This study [12], explores the benefits of employing a DCNN for SER. A pre-trained network is utilized to extract features from speech emotion datasets. Additionally, a CFS technique is applied to identify the most pertinent and discriminative features for SER. Classification is performed using support vector machines, random forests, k-nearest neighbors, and neural network classifiers. Experiments are conducted on four publicly available datasets. The proposed method achieves accuracies, 95.10%, 82.10%, 83.80%, and 81.30%, Emo-DB, SAVEE, IEMOCAP, and RAVDESS in speaker-dependent SER experiments, respectively. Also achieves accuracies, 90.50%, 66.90%, 76.60%, and 73.50%, Emo-DB, SAVEE, IEMOCAP, and RAVDESS in speaker-independent SER experiments, respectively. In this study [13], Emotion classification was performed using a SVM, and DA is achieved using a generative adversarial network (GAN). Two FS, Fisher and LDA, are simultaneously employed in the GAN to identify relevant and informative features while eliminating redundant and irrelevant ones. Researchers were conducted the experiments using Python on four widely used databases, considering emotions such as sadness, fear, anger, happiness, and neutral. The proposed method's results are compared with classical DR techniques. The proposed approach achieved accuracies of 86.32%, 59.77%, 73.82%, and 61.67% for the EMO-DB, eNTERFACE05, SAVEE, and IEMOCAP datasets, respectively. In comparison, classical dimension reduction techniques yield accuracies of 85.20% for 1582 main features, 86.00% for 1582 main features+added features, 81.71% for PCA, and 86.32% for The proposed method. This research [14], introduces a non-linear signal quantification approach based on entropy features, specifically the randomness measure. Initially, the speech signals are decomposed into Intrinsic Mode Functions (IMFs), and these IMFs are further divided into dominant frequency bands: high frequency, mid-frequency, and base frequency. A feature vector is constructed using the computed entropy measures, incorporating the randomness feature for all emotional signals. State-of-the-art classifiers, including LDA, Naïve Bayes, K-NN, SVM, RF, and Gradient Boosting Machine, are trained using the feature vector. The proposed method achieves promising results through a tenfold cross-validation on the Toronto Emotional Speech dataset, with the LDA classifier demonstrating a peak balanced accuracy of 93.3%, an F1 score of 87.9%, and an area under the curve (AUC) value of 0.995 in recognizing emotions from speech signals of native English speakers. This study [24] to improve the accuracy of emotion recognition systems, a combination of CNN and LSTM studied and used. This ConvLSTM model effectively captures crucial audio features, leading to enhanced performance in SER tasks. D) techniques were employed to augment the training data, while feature selection using the Analysis of Variance (ANOVA) technique was utilized to identify informative features. The proposed approach demonstrates its effectiveness in identifying optimal configurations, resulting in high accuracy rates for emotion classification. The obtained results showcase the capability of the proposed approach in accurately classifying speech emotions across diverse datasets. Specifically, accuracy rates of 95.34% for RAVDESS, 96.03% for SAVEE, 97.00% for EMO-DB, and 99.75% for TESS were achieved. Combining the RAVDESS, SAVEE, and EMODB datasets resulted in an accuracy of 97.37% for R+S+E, while combining RAVDESS, SAVEE, and TESS datasets resulted in an accuracy of 98.94% for R+S+T.

Material and methods

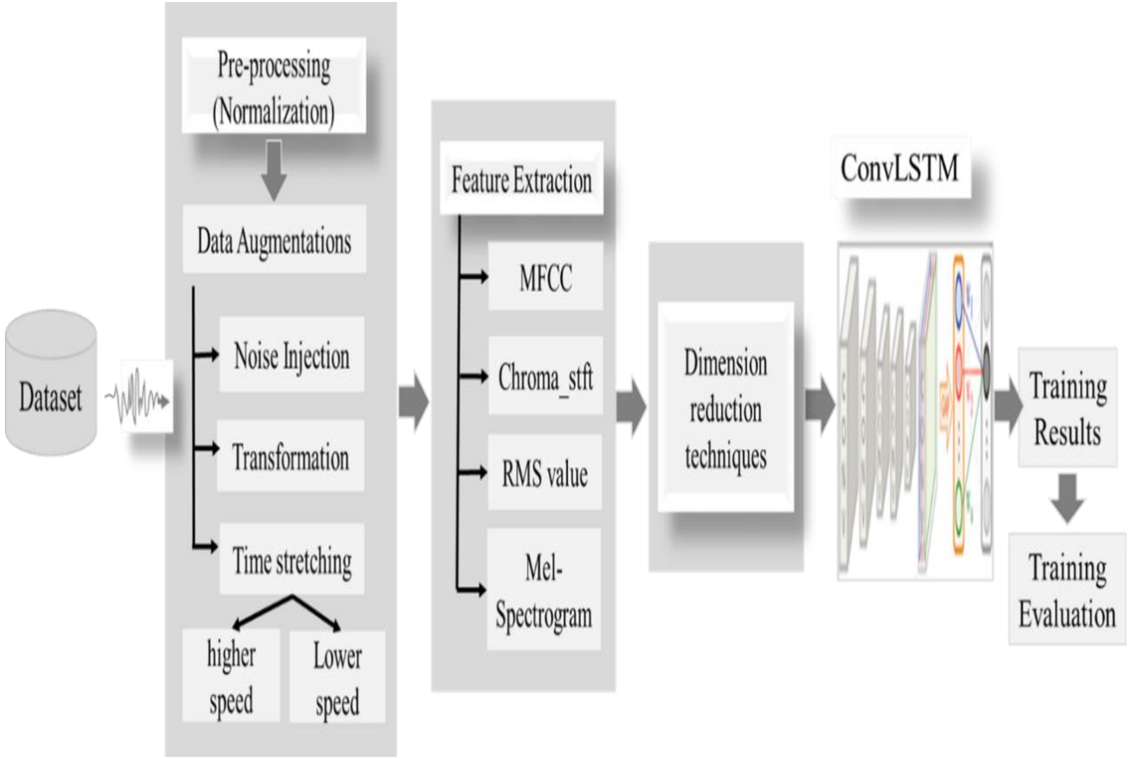


Figure 2: The proposed methodology.

This research study introduces a ConvLSTM approach, as illustrated in Figure 2.The figure presents schematic representations of the sequential steps involved in this approach. Subsequently, this approach is comprehensively elaborated upon in this section, providing a detailed description of its methodology and implementation.

Datasets Description

In this study, three distinct audio datasets were utilized, namely RAVDESS, SAVEE, and EMO-DB. These datasets are widely employed by researchers in the field of emotion recognition.

- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):

The RAVDESS dataset is designed for emotion recognition and provides a comprehensive collection of audio visual recordings. It includes recordings from 24 professional actors, with an equal split of 12 males and 12 females, each participating in 60 trials, resulting in a total of 1440 data samples. The dataset covers a wide range of emotions, such as calmness, happiness, sadness, anger, fear, and surprise. Each actor performed scripted sentences and vocalizations while portraying different emotional states. RAVDESS is commonly utilized for developing emotion recognition models, studying audio visual processing, and analysing emotional speech and singing [15].

- SAVEE (Surrey Audio-Visual Expressed Emotion):
  SAVEE is a dataset specifically created to facilitate the recognition of spoken emotions. It consists of audio recordings from four male actors who express seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutral. The dataset contains over 480 British English utterances, where each actor delivers emotions through scripted sentences. SAVEE is commonly employed in research related to emotion recognition, speech processing, and machine learning algorithms for emotion classification [16].
- EMO-DB (Emotional Database):
  EMO-DB is a recognized dataset primarily focused on emotional speech recognition. It comprises audio recordings from ten German actors, with an equal representation of male and female speakers. EMO-DB comprises 535 German audio utterances categorized into seven emotion classes including happiness, sadness, anger, fear, and boredom. EMO-DB is extensively used in research domains such as emotion recognition, speech analysis, and affective computing, providing valuable resources for studying emotional expression in speech and developing emotion [17].

Data Pre-Processing

To preprocess the audio recordings in this study, the librosa library was employed in Python. Librosa is a powerful library specifically designed for audio processing tasks and feature extraction. It offers a wide range of functions and utilities to convert audio recordings into a suitable digital representation for further analysis.

Additionally, a normalization technique was employed to enhance the quality and consistency of the data as shown in Figure 3. Normalization helps to scale the audio signals to a standardized range, reducing variations in amplitude and ensuring that all samples have a comparable dynamic range. This step is essential for achieving better performance and avoiding biases caused by differences in signal magnitude across different recordings [18].

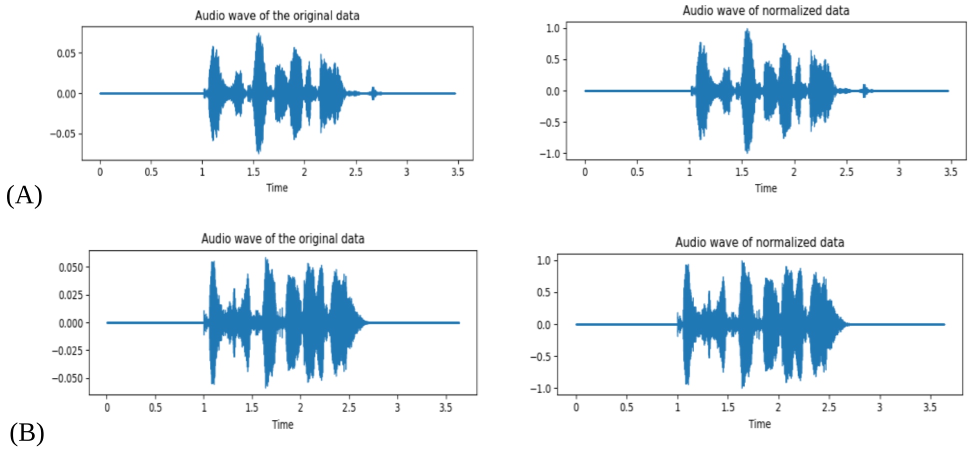


Figure 3: (A) Audio wave for emotion of fear. (B) Audio wave for emotion of happiness

By leveraging the librosa library for audio conversion and incorporating a normalization technique, the audio data in this study were preprocessed effectively, enabling subsequent feature extraction and facilitating accurate speaker emotion recognition algorithms.

Data Augmentation

Data augmentation is a widely used technique in Speaker Emotion Recognition (SER) models to enhance their robustness, generalization, and mitigate overfitting. By applying various transformations to the training data, this technique expands the dataset's size and diversity. This, in turn, improves the model's accuracy and promotes data distribution invariance by introducing variations that mimic real-world scenarios. To address the data imbalance among emotion classes, several techniques were implemented to augment the sample size within the dataset [19]. These techniques include:

1. Noise Injection: Random noise is added to the audio data by incorporating a noise rate of 0.035. This introduces additional variations to the audio.

2. Time Stretching: This approach extends or compresses the time series of the audio at a constant rate. The rate parameters used in this study are set to 0.5 and 1.25. A rate of 0.5 results in slower audio, while a rate of 1.25 produces faster audio.
3. Time Shifting: This technique randomly shifts the audio to the left or right by a random duration within the range of -9 to 9 seconds. When shifting forward, the initial seconds are set to 0, while shifting backward maintains the final seconds as 0.

By employing these data augmentation techniques as shown in Figure 4 , the study aimed to increase the dataset's diversity, alleviate data imbalance, and enhance the performance of the Speaker Emotion Recognition models.

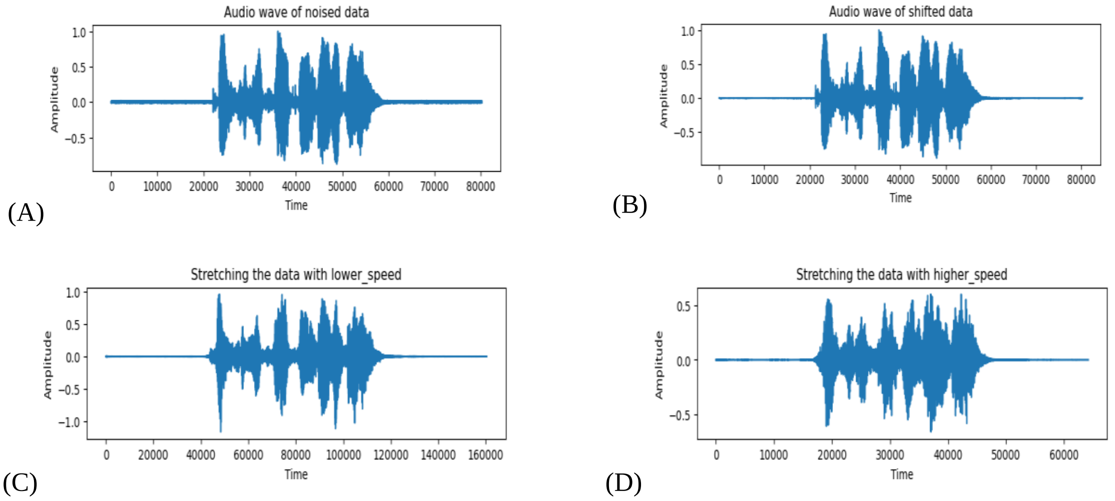


Figure 4: The signals of augmentation of audio signals (A) Audio signal with noised data; (B) Audio signal with shifted data; (C) stretching the signal with lower speed; (D) Stretching the signal with higher speed.

Feature Extraction

Accurately extracting important features from speech audio signals is crucial for SER tasks and improving model accuracy. In this study, the proposed dilated ConvLSTM model utilize four spectral features: MFCC, Mel-spectrogram, Chromagram, and RMS values as shown in Figure 5. These features are briefly described as follows:

MFCC (Mel-frequency cepstral coefficients): MFCCs capture vocal tract characteristics essential for distinguishing emotions in speech. They involve framing, windowing, discrete Fourier transform (DFT), logarithm of magnitude, Mel frequency warping, and discrete cosine transform (DCT). MFCCs provide a concise representation of spectral features related to the vocal tract, enabling accurate emotion detection in Human-Computer Interaction (HCI) systems [20].

Mel-Spectrogram: Mel spectrogram is a representation that captures energy distribution in audio signals across frequency and time. It improves the audio spectrum representation based on the human ear's response pattern. It involves dividing the signal into frames, applying a window function, and performing Fourier transform. The obtained spectrum is then filtered using a Mel filter array and subjected to log compression, resulting in a logarithmic spectral representation of the audio signal [21].

Chromagram: The chroma feature captures tonal characteristics in audio signals. It partitions the spectrum into frequency bins and assigns them to specific pitch classes based on proximity to corresponding pitches in a chromatic scale. Energy within each bin is weighted to reflect its association with the pitch class. The resulting 12-dimensional vector characterizes the presence or intensity of each pitch class in the audio signal.

Root Mean Square (RMS): RMS quantifies the power content and loudness of speech signals at the frame level. Including RMS as a feature enhances emotion classification accuracy, as different emotions exhibit variations in intensity. It is computed by taking the square root of the average squared amplitudes, capturing the energy characteristics of speech signals and contributing to more accurate SER models [22].

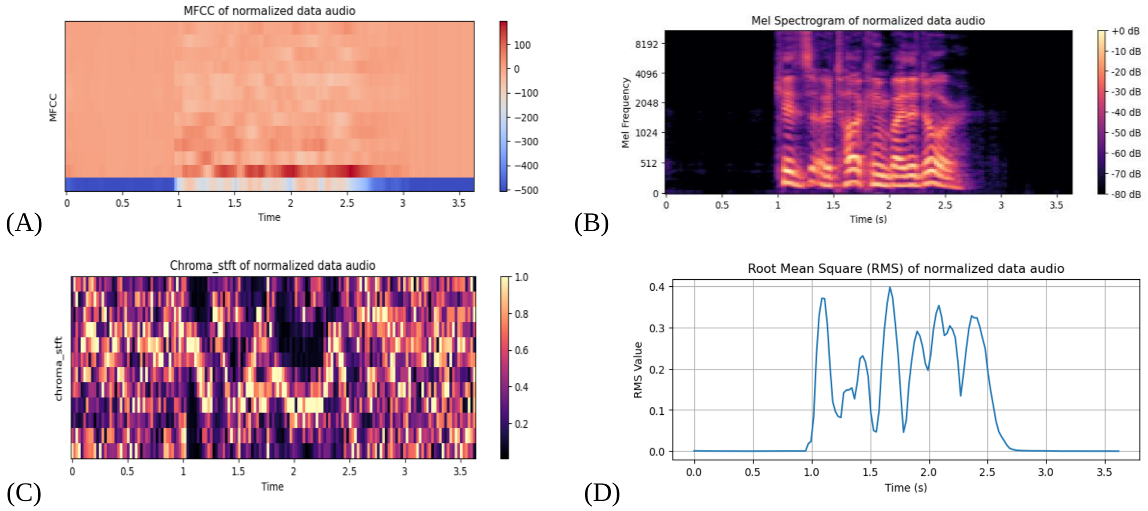


Figure 5: The features that extracted from each data sample.

Dimensionality Reduction

In the dimension reduction stage, two FS techniques, which are CFS and ANOVA, were applied. In addition, FR techniques: PCA and LDA used to reduce the dimensionality of the data.

The CFS technique evaluates the relationships between features by measuring their correlations. It aims to select a subset of features that are highly correlated with the target variable while minimizing redundancy among them. By identifying and retaining the most relevant features, CFS helps in reducing dimensionality while preserving the discriminatory power of the data [12].

The ANOVA test, on the other hand, assesses the statistical significance of feature variations across different classes or groups. It measures the differences in means between groups and determines whether these differences are significant. By selecting features with high variance between classes, the ANOVA test helps in identifying features that are most informative and contribute significantly to the classification or regression tasks [23].

The PCA, is a widely used technique that aims to transform the original features into a new set of uncorrelated variables called principal components. These components are ordered by the amount of variance they explain in the data. By retaining a subset of the most important principal components, PCA allows for significant dimension reduction while preserving the maximum amount of information contained in the data [6].

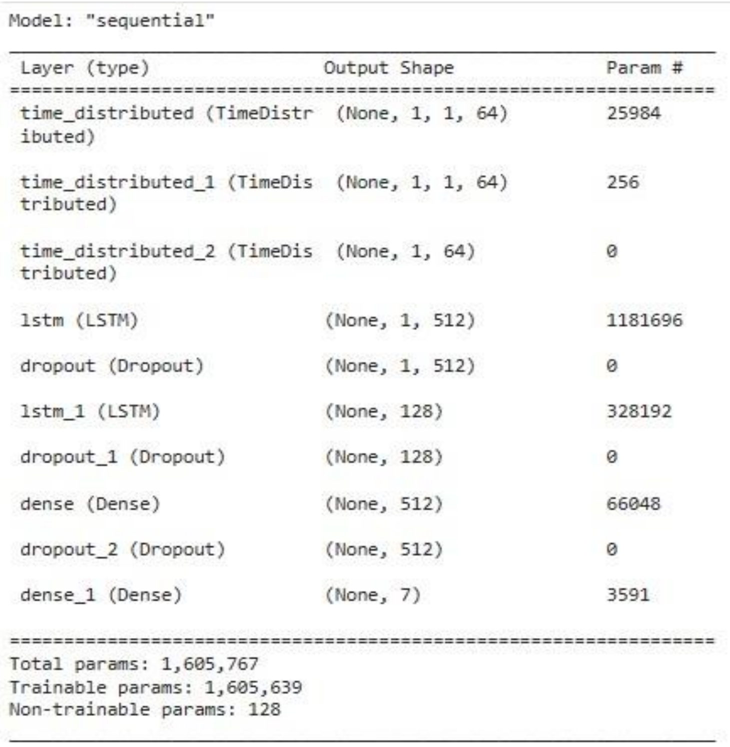
The LDA, on the other hand, focuses on finding a linear projection that maximizes the separability between different classes or groups in the data. It aims to find a lower-dimensional subspace where the between-class scatter is maximized, while the within-class scatter is minimized. By projecting the data onto this subspace, LDA facilitates dimension reduction while maximizing the class discriminability, making it suitable for classification tasks [6].

By employing these feature selection and feature reduction techniques, the dimensionality of the data was effectively reduced, allowing for a more concise representation of the original data while retaining the discriminative and informative aspects.

Model Architecture

This study introduces a ConvLSTM model for handling sequential time-series data in a categorical classification problem. The preprocessing techniques applied to the audio signals played a crucial role in enhancing the model's learning capability and generalization performance. Various preprocessing steps, including noise injection, time shifting, and time stretching, were applied to optimize the quality of the input data. Additionally, pertinent acoustic features such as MFCC, Mel-spectrogram, Chroma gram, and RMS were extracted to effectively capture and represent the relevant acoustic characteristics. Dimension reduction techniques were further employed for feature selection to refine the input representation by retaining informative and discriminatory features.

The ConvLSTM model addresses the challenges of vanishing gradients and over-reliance on recent states in sequential time-series data. It consists of sequential layers that extract features and transform the data into binary classifiers. An LSTM layer with 512 units handles the sequential time-series data, and dropout layers are incorporated to prevent overfitting. Additional LSTM and dense layers are included for capturing patterns and conducting feature extraction. The final dense layer, with 7 or 8 units depending on the dataset, utilizes SoftMax activation for accurate classification into distinct categories. The integration of these layers enables the effective processing of sequential data and achieves precise classification. The model architecture and layer structure are visually represented in Figure 6.



```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 time_distributed (TimeDistr  (None, 1, 1, 64)         25984
 ibuted)

 time_distributed_1 (TimeDis  (None, 1, 1, 64)         256
 tributed)

 time_distributed_2 (TimeDis  (None, 1, 64)            0
 tributed)

 lstm (LSTM)                 (None, 1, 512)            1181696

 dropout (Dropout)           (None, 1, 512)            0

 lstm_1 (LSTM)               (None, 128)               328192

 dropout_1 (Dropout)         (None, 128)               0

 dense (Dense)               (None, 512)               66048

 dropout_2 (Dropout)         (None, 512)               0

 dense_1 (Dense)             (None, 7)                 3591

=================================================================
Total params: 1,605,767
Trainable params: 1,605,639
Non-trainable params: 128
_________________________________________________________________
```

Figure 6: ConvLSTM model structure.

Explement Setup

The experiment utilized the Jupyter Notebook environment with Python version 3.9.18. Data augmentation and feature extraction processes were performed using an Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz 2.30 GHz processor. The proposed model made use of the NVIDIA GeForce MX130 graphics card. Implementation of the framework involved the scikit-learn Python library for machine learning, along with other supportive libraries.

In order to ensure a robust evaluation and facilitate comparisons with previous studies, the dataset was approached in two distinct manners. Firstly, an 80:20% train-test split was employed, allowing the model to learn from the majority of the data while validating its performance on unseen samples. Secondly, cross-validation was applied using 4-fold to further scrutinize the model's performance and ascertain its generalization capabilities.

Experiments and Results

This experiment aimed to investigate the impact of dimensionality reduction techniques on model accuracy using diverse datasets in the context of classification tasks.

In the first experiment, utilizing an 80:20% training-test split and applying data augmentation techniques, the CFS feature selection method achieved high accuracy rates. The RAVDESS dataset achieved an accuracy rate of 97.22%, SAVEE achieved 96.65%, and EMODE achieved 97.75%. Cross-validation was also conducted, resulting in even higher accuracy.

In the second experiment, the ANOVA test technique replaced the CFS feature selection method. The accuracy rates achieved were slightly lower compared to the first experiment. RAVDESS achieved an accuracy rate of 97.01%, SAVEE achieved 97.70%, and EMODE achieved 97.19%.

The third experiment utilized the PCA dimensionality reduction technique. Although the accuracy rates were slightly lower compared to the first two experiments, they were still considerable. RAVDESS achieved 93.88%, SAVEE achieved 95.19%, and EMODE achieved 95.70%.

The fourth experiment employed the LDA dimensionality reduction technique. The accuracy rates achieved were significantly lower compared to the previous experiments. RAVDESS achieved 73.75%, SAVEE achieved 88.93%, and EMODE achieved 92.33%.

The application of feature selection and reduction techniques in this study successfully decreased the dimensionality of the data, resulting in a more compact representation that preserved the significant and informative characteristics. Through the experiments, it was observed that the model's accuracy improved notably when a subset of 135 features was selected. The results were shown the performance of different feature selection and feature reduction techniques on three datasets: RAVDESS, SAVEE, and EMO-DB, as shown in Table1.

Table1: Accuracy (%) achieved by using different Dimension reduction techniques.

Dataset	Train-Test split (80:20%)				4 Fold			
	FS		FR		FS		FR	
	CFS	ANOVA	PCA	LDA	CFS	ANOVA	PCA	LDA
RAVDESS	97.22%	97.01%	93.88%	73.75%	97.11 %	96.95 %	95.50%	74.18 %
SAVEE	96.65%	97.70%	95.19%	88.93%	96.20%	96.65%	93.90%	86.72 %
EMO-DB	97.75%	97.19%	95.70 %	92.33%	98.39%	98.61%	97.57%	90.46%

Comparative Analysis

Comparative Analysis Between Different Dimension Reduction Techniques.

This section presents a comparative analysis of the results obtained after applying the proposed methodology to three distinct datasets for evaluating the performance of dimensionality reduction techniques. The performance differences between these techniques are illustrated in the graph shown in Figure 7. By comparing these results, the aim to gain insights into the effectiveness of these techniques and their suitability for different datasets and classification tasks. This comparative analysis will contribute to a better understanding of dimension reduction methods and provide guidance in selecting the most appropriate technique to optimize classification performance.

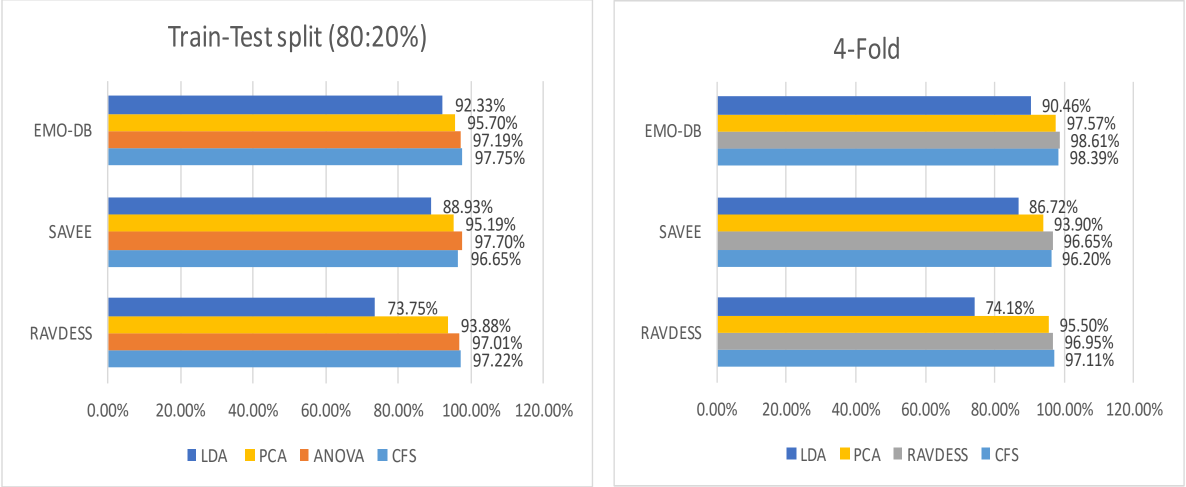


Figure 7: Performance (Accuracy%) comparison of three datasets with Dimension reduction techniques.

- Following the application of the proposed methodology on three distinct datasets for performance evaluation and learning assessment, for the feature selection techniques, CFS consistently yielded the highest accuracy rates across all datasets and evaluation methods. In the 80:20% train-test split, CFS achieved accuracy rates of 97.22% for RAVDESS, 96.65% for SAVEE, and 97.75% for EMO-DB. Similarly, in the 4-fold cross-validation, CFS achieved high accuracy rates ranging from 97.11% to 98.39%. This indicates that CFS effectively selected informative features and contributed to accurate classification.
- On the other hand, the ANOVA test technique also performed well, but slightly lower compared to CFS by 0.21%. It achieved accuracy rates ranging from 97.01% to 97.70% in the training test section and from 96.95% to 98.61% in four-fold cross-validation. The ANOVA test proved capable of identifying significant differences between categories, which contributed to accurate classification. The ANOVA technique outperformed CFS in some data sets, and in others CFS outperformed ANOVA as shown in Table 1. It is worth noting that it was noted that differences in data sets in terms of data balance, languages, pronunciation, and dialects affect accuracy of dimension reduction techniques.
- Regarding feature reduction techniques, PCA and LDA were compared. PCA, despite reducing the dimensionality of the data, achieved lower accuracy rates compared to the feature selection techniques. It achieved accuracy rates ranging from 93.88% to 95.70% in the train-test split and from 95.50% to 97.57% in the 4-fold cross-validation. Although PCA reduced the dimensionality, it might have discarded some discriminative information, resulting in a slight decrease in accuracy.
- In contrast, LDA, which also aimed at dimensionality reduction, yielded significantly lower accuracy rates compared to other techniques. It achieved accuracy rates ranging from 73.75% to 92.33% in the train-test

split and from 74.18% to 90.46% in the 4-fold cross-validation. This suggests that LDA might not have captured the discriminative structure of the data effectively, leading to a decrease in classification accuracy.

Overall, the results indicate that CFS was the most effective feature selection technique, while ANOVA test performed well too. For feature reduction, PCA was relatively effective, while LDA resulted in lower accuracy rates Figure 8) shows the difference in accuracy between CFS and other dimensional techniques. These findings emphasize the importance of careful selection and evaluation of feature selection and reduction techniques based on the specific dataset and classification task at hand.

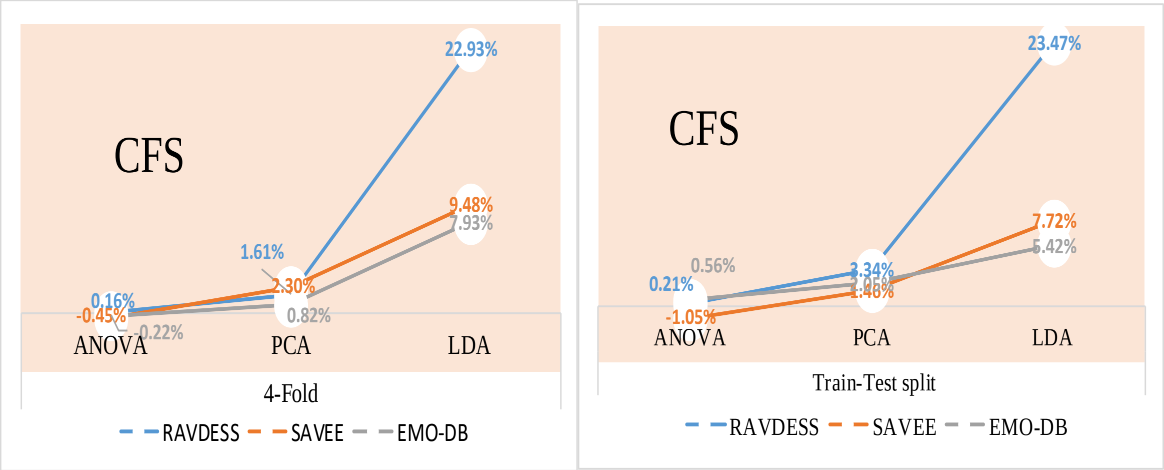


Figure 8: Accuracy Differences between CFS and Other Dimensional Techniques.

The obtained results were compared and analyzed through the utilization of performance evaluation metrics such as accuracy rate, loss, rating report, and confusion matrix. Figure 9 depicts the results of the performance of the proposed model using CFS technology. By examining these metrics, the experiment aimed to gain insights into the effectiveness of dimensionality reduction techniques and their influence on model performance across different datasets. In this study it are presented results of the performance of the proposed model using CFS technology for datasets divided into an 80:20% train-test split as shown in Table 2 to Figure 5.

Table 2: Performance of the proposed model using CFS technology on datasets that partitioned 80:20%.

Emotion	RAVDESS			Emotion	EMO-DB			Emotion	SAVEE		
	precision	recall	f1-score		precision	recall	f1-score		precision	recall	f1-score
angry	0.99	0.99	0.99	angry	1.00	1.00	1.00	angry	1.00	1.00	1.00
calm	0.98	0.99	0.99	boredom	1.00	0.92	0.96	disgust	1.00	0.92	0.96
disgust	0.96	0.99	0.97	disgust	1.00	0.92	0.96	fear	1.00	0.92	0.96
fear	0.98	0.95	0.97	fear	1.00	1.00	1.00	happy	1.00	1.00	1.00
happy	0.97	0.97	0.97	happy	0.95	1.00	0.98	neutral	0.95	1.00	0.98
neutral	0.99	0.93	0.96	neutral	0.96	0.99	0.97	sad	0.96	0.99	0.97
sad	0.98	0.96	0.97	sad	0.89	1.00	0.94	surprise	0.89	1.00	0.94
surprise	0.93	0.97	0.95	-	-	-	-	-	-	-	-

Table 3: Performance of the proposed model using ANOVA technology on datasets.

Emotion	RAVDESS			Emotion	EMO-DB			Emotion	SAVEE		
	precision	recall	f1-score		precision	recall	f1-score		precision	recall	f1-score
angry	0.99	0.97	0.98	angry	0.99	0.99	0.99	angry	1.00	0.97	0.98
calm	0.98	0.97	0.98	boredom	1.00	0.91	0.95	disgust	1.00	0.97	0.98

disgust	0.96	0.97	0.96	disgust	0.96	0.96	0.96	fear	0.98	0.98	0.98
fear	0.96	0.97	0.97	fear	0.97	1.00	0.99	happy	0.97	0.98	0.98
happy	0.97	0.97	0.97	happy	0.97	0.97	0.97	neutral	0.96	0.99	0.98
neutral	0.99	0.92	0.95	neutral	0.91	0.97	0.94	sad	0.97	0.96	0.96
sad	0.95	0.98	0.96	sad	0.98	1.00	0.99	surprise	0.98	0.98	0.98
surprise	0.97	0.98	0.98	-	-	-	-	-	-	-	-

Table 4: Performance of the proposed model using PCA technology on datasets.

Emotion	RAVDESS			Emotion	EMO-DB			Emotion	SAVEE		
	precision	recall	f1-score		precision	recall	f1-score		precision	recall	f1-score
angry	0.97	0.96	0.96	angry	0.96	1.00	0.98	angry	1.00	0.93	0.97
calm	0.99	0.94	0.97	boredom	0.99	0.88	0.95	disgust	0.93	1.00	0.96
disgust	0.93	0.95	0.94	disgust	1.00	1.00	1.00	fear	0.93	0.89	0.91
fear	0.96	0.90	0.93	fear	0.99	0.96	0.97	happy	0.94	0.98	0.96
happy	0.94	0.93	0.93	happy	1.00	0.94	0.97	neutral	0.98	0.95	0.97
neutral	0.92	0.89	0.90	neutral	0.86	0.99	1.00	sad	0.96	0.97	0.96
sad	0.89	0.94	0.92	sad	1.00	1.00	0.99	surprise	0.89	0.91	0.90
surprise	0.90	0.97	0.93	-	-	-	-	-	-	-	-

Table 5: Performance of the proposed model using LDA technology on datasets.

Emotion	RAVDESS			Emotion	EMO-DB			Emotion	SAVEE		
	precision	recall	f1-score		precision	recall	f1-score		precision	recall	f1-score
angry	0.86	0.79	0.82	angry	0.93	0.95	0.94	angry	0.88	0.93	0.9
calm	0.79	0.81	0.80	boredom	0.92	0.97	0.94	disgust	0.86	0.86	0.86
disgust	0.74	0.67	0.70	disgust	0.91	0.96	0.94	fear	0.8	0.91	0.85
fear	0.78	0.71	0.74	fear	0.93	0.89	0.91	happy	0.93	0.89	0.91
happy	0.67	0.68	0.68	happy	0.87	0.85	0.85	neutral	0.94	0.91	0.92
neutral	0.75	0.64	0.69	neutral	0.93	0.84	0.89	sad	0.89	0.96	0.92
sad	0.63	0.71	0.66	sad	0.98	1.00	0.99	surprise	0.85	0.75	0.8
surprise	0.71	0.84	0.77	-	-	-	-	-	-	-	-

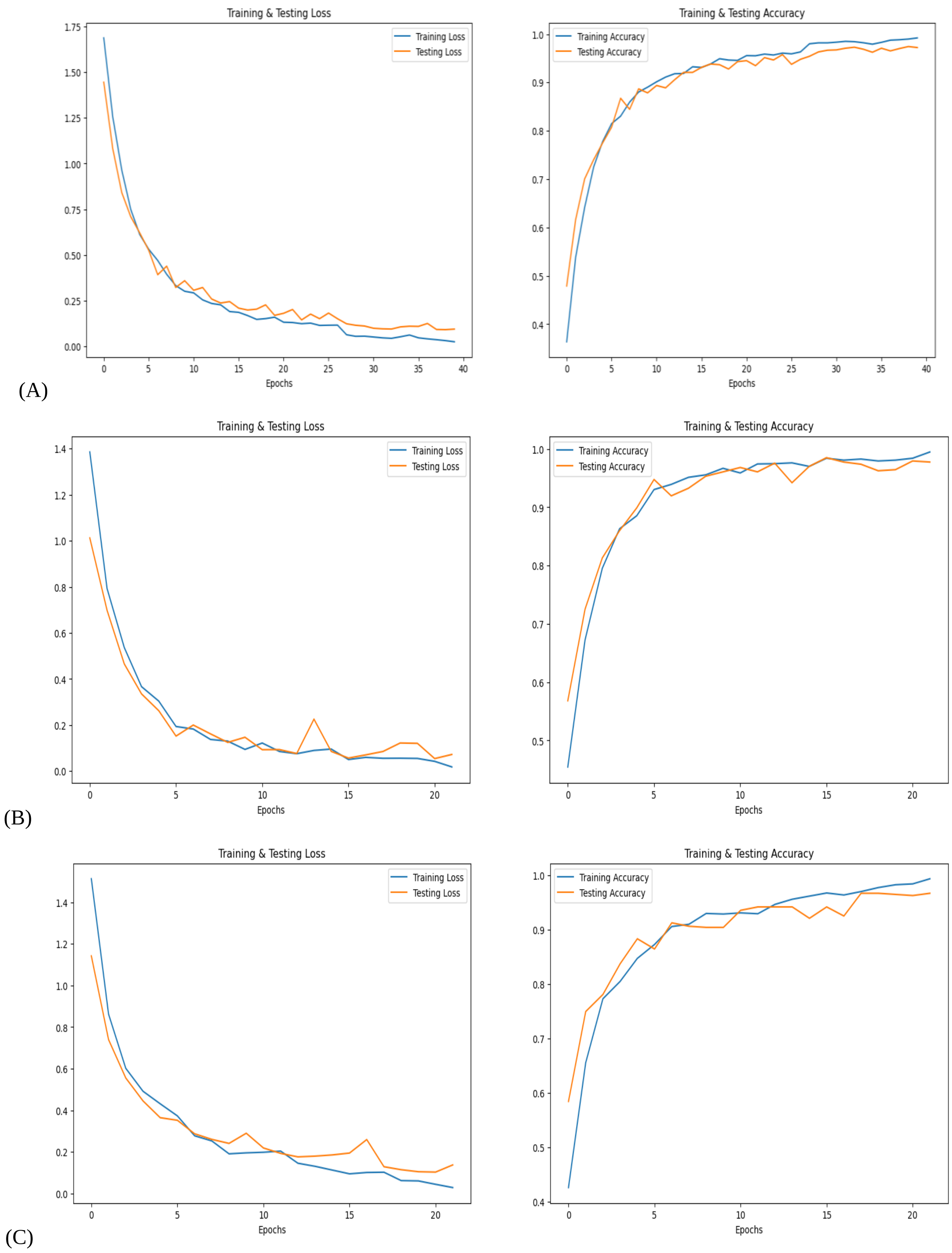


Figure 9: The accuracy and the loss value of the best result for each dataset (A) RAVDESS, (B) EMO-DB and (C) SAVEE with CFS.

Comparative Analysis with State-of-the-Art Works

This section presents a comparative analysis of the obtained results in relation to prior research documented in the literature, with a focus on evaluating the impact of dimensionality reduction techniques on classification accuracy. The comparison is based on the average accuracy achieved by dividing the datasets into an 80:20% split. The results demonstrate the superiority of the proposed approach over all the studies considered in this investigation. The comparison clearly indicates that the selection of feature selection techniques has a significant influence on the accuracy of emotion recognition. Several recently published works were compared with the findings of this study. The results reveal that the proposed approach achieved a 3.61% improvement in accuracy for the RAVDESS dataset when employing the CFS technique. Similarly, for the SAVEE dataset, there was a notable improvement of 7.88% compared to competing methods. Additionally, the proposed approach yielded a 5.73% improvement in accuracy for the EMO-DB dataset. Overall, the results highlight the superiority of the proposed approach over the existing studies in this field. The comparison underscores the influential role of feature selection techniques in enhancing the accuracy of emotion recognition as shown as Table 6).

Table 6: Comparison of accuracy experiments of the proposed CONVLSTM approach with existing methods.

Study	Dataset	The proposed methods in studies	Accuracy of the proposed models in studies	Accuracy of models with one of the selected DR techniques			
				CFS	ANOVA	PCA	LDA
[9]	EMO-DB	AlexNet + CFS + MLP	-	92.02%	-	-	-
	SAVEE	AlexNet + CFS + SVM	-	88.77%	-	-	-
	RAVDESS	AlexNet + CFS + SVM	-	93.61%	-	-	-
[11]	EMO-DB	pQPSO	82.82%	-	-	66.16%	67.36%
	SAVEE	pQPSO	60.79%	-	-	51.47%	50.49%
[12]	EMO-DB	DCNN	-	90.50%	-	-	-
	SAVEE	DCNN	-	66.90%	-	-	-
	RAVDESS	DCNN	-	73.50%	-	-	-
[13]	EMO-DB	A hybrid method	86.32	-	-	81.71	-
	SAVEE	A hybrid method	73.82	-	-	72.39	-
[24]	EMO-DB	ConvLSTM	-	-	97.00%	-	-
	SAVEE	ConvLSTM	-	-	96.03%	-	-
	RAVDESS	ConvLSTM	-	-	95.34%	-	-
The proposed	RAVDESS	ConvLSTM	-	97.22%	97.01%	93.88%	73.75%
	SAVEE	ConvLSTM	-	96.65%	97.70%	95.19%	88.93%
	EMO-DB	ConvLSTM	-	97.75%	97.19%	95.70%	92.33%

Conclusion and Future Work

The research highlights the significance of FS and FR techniques in SER systems. The study focuses on the importance of DR algorithms in the SER methodology and investigates the impact of four DR techniques on model learning accuracy and generalization. A combined approach was used CNN and LSTM to propose the ConvLSTM model, aiming to enhance the accuracy of speech emotion recognition, by improving the quality of the input data by using four DA techniques and extracting relevant acoustic features such as MFCC, Mel-spectrogram, Chromagram, and RMS. After then, the CFS, ANOVA, PCA, and LDA techniques were utilized to select the most useful and discriminating features. Experiments were conducted to illustrate the impact of DR on the evaluation of the model was conducted using RAVDESS, SAVEE, and EMO-DB datasets. Based on the results of these experiments, the FS techniques to be the best for feature selection, as the experiments demonstrated the effectiveness of the CFS method, that achieved high accuracy rates across all datasets. Additionally, the PCA technique performed well for FR. The comparative analysis shows that the proposed approach achieved significant accuracy improvements, ranging from 3.61% to 7.88%, compared to existing methods. Which indicates that the correct selection of feature selection techniques contributes significantly to improving accuracy and generalization. Through these experiments, it was found that choosing the appropriate dimension reduction technique depends on several factors, the dataset type, data size, compatibility with subsequent models, It was observed that during the experiments, the CFS technique outperformed ANOVA in two datasets, and the ANOVA technique outperformed CFS in one dataset.

Future work in the field of this study could involve investigating the effectiveness of combining different feature selection (FS) and dimensionality reduction (DR) techniques to achieve optimal feature representation and dimensionality reduction for speech emotion recognition (SER), leading to potential enhancements in accuracy and generalization. Additionally, exploring the analysis of more diverse datasets that encompass a wide range of languages, dialects, cultural backgrounds, and emotional expressions would provide a better understanding of how FS and DR techniques perform across different contextual settings. These research endeavors would contribute to advancing the field of SER by improving accuracy, expanding applicability, and gaining insights into the performance of feature selection and dimensionality reduction techniques in diverse scenarios.

References

[1] J. Lowgren, J. M. Carroll, M. Hassenzahl, T. Erickson, and A. Blackwell, “The Encyclopedia of Human-Computer Interaction,” *Interaction design foundation*, 2019.

[2] A. A. Alnuaim *et al.*, “Human-Computer Interaction with Detection of Speaker Emotions Using Convolution Neural Networks,” *Computational Intelligence and Neuroscience*, vol. 2022, p. e7463091, Mar. 2022, doi: 10.1155/2022/7463091.

[3] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, “Bagged support vector machines for emotion recognition from speech,” *Knowledge-Based Systems*, vol. 184, p. 104886, Nov. 2019, doi: 10.1016/j.knosys.2019.104886.

[4] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, “Deep learning applications and challenges in big data analytics,” *Journal of Big Data*, vol. 2, no. 1, p. 1, Feb. 2015, doi: 10.1186/s40537-014-0007-7.

[5] W. Jia, M. Sun, J. Lian, and S. Hou, “Feature dimensionality reduction: a review,” *Complex Intell. Syst.*, vol. 8, no. 3, pp. 2663–2693, Jun. 2022, doi: 10.1007/s40747-021-00637-x.

[6] S. Nanga *et al.*, “Review of Dimension Reduction Methods,” *Journal of Data Analysis and Information Processing*, vol. 9, no. 3, Art. no. 3, Jul. 2021, doi: 10.4236/jdaip.2021.93013.

[7] X. Huang, L. Wu, and Y. Ye, “A Review on Dimensionality Reduction Techniques,” *Int. J. Patt. Recogn. Artif. Intell.*, vol. 33, no. 10, p. 1950017, Sep. 2019, doi: 10.1142/S0218001419500174.

[8] M. BÜYÜKKEÇECİ and M. Okur, “A Comprehensive Review of Feature Selection and Feature Selection Stability in Machine Learning,” *GAZI UNIVERSITY JOURNAL OF SCIENCE*, vol. 36, Sep. 2022, doi: 10.35378/gujs.993763.

[9] A. Amjad, L. Khan, and H.-T. Chang, “Effect on speech emotion classification of a feature selection approach using a convolutional neural network,” *PeerJ Computer Science*, vol. 7, p. e766, Nov. 2021, doi: 10.7717/peerj-cs.766.

[10] R. Aziz, C. K. Verma, and N. Srivastava, “Dimension reduction methods for microarray data: a review,” *AIMS Bioengineering*, vol. 4, pp. 179–197, Mar. 2017, doi: 10.3934/bioeng.2017.2.179.

[11] F. Daneshfar and S. Kabudian, “Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm,” *Multimedia Tools and Applications*, vol. 79, Jan. 2020, doi: 10.1007/s11042-019-08222-8.

[12] M. Farooq, F. Hussain, N. K. Baloch, F. Raja, H. Yu, and Y. Zikria, “Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network,” *Sensors*, vol. 20, Oct. 2020, doi: 10.3390/s20216008.

[13] A. Shilandari, H. Marvi, and N. Hadjiabdolhamid, “Effective Feature Selection in Speech Emotion Recognition Systems using Generative Adversarial Networks.” Nov. 11, 2022. doi: 10.21203/rs.3.rs-2244414/v1.

[14] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, “Emotion classification from speech signal based on empirical mode decomposition and non-linear features,” *Complex Intell. Syst.*, vol. 7, no. 4, pp. 1919–1934, Aug. 2021, doi: 10.1007/s40747-021-00295-z.

[15] L. Sr and R. Fa, “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, May 2018, doi: 10.1371/journal.pone.0196391.

[16] P. Jackson and S. ul haq, “Surrey Audio-Visual Expressed Emotion (SAVEE) database.” Apr. 01, 2011.

[17] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, *A database of German emotional speech*, vol. 5. 2005, p. 1520. doi: 10.21437/Interspeech.2005-446.

[18] D. Singh and B. Singh, “Investigating the impact of data normalization on classification performance,” *Applied Soft Computing*, p. 105524, May 2019, doi: 10.1016/j.asoc.2019.105524.

[19] C. Alves *et al.*, “Transfer Learning and Data Augmentation Techniques applied to Speech Emotion Recognition in SE&R 2022,” 2022, [Online]. Available: https://github.com/BrunoGianesi/Speaker-Gender-Recognition.

[20] S. P. Mishra, P. Warule, and S. Deb, “Speech emotion recognition using MFCC-based entropy feature,” *SIViP*, Aug. 2023, doi: 10.1007/s11760-023-02716-7.

[21] S. S. Stevens, J. Volkmann, and E. B. Newman, “A Scale for the Measurement of the Psychological Magnitude Pitch,” *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, Jun. 2005, doi: 10.1121/1.1915893.

[22] Md. Rayhan Ahmed, S. Islam, A. K. M. Muzahidul Islam, and S. Shatabda, “An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition,” *Expert Systems with Applications*, vol. 218, p. 119633, May 2023, doi: 10.1016/j.eswa.2023.119633.

[23] W. Zhang and Y. Qi, “ANOVA-nSTAT: ANOVA methodology and computational tools in the paradigm of new statistics,” vol. 14, pp. 48–67, Mar. 2024.

[24] R. M. Ben-Sauod, R, S, Alshwihde, W, I, Eltarhouni, “ENHANCING Speech Emotion Recognition through a Cross-Dataset Analysis: Exploring Improved Models,” IEEE Xplore., in press.