An

Internship Training Report

Entitled

Speech Emotion Recognition with Combination of CNN-LSTM and DNN

Submitted to the Department of Electronics Engineering in Partial Fulfilment for the Requirements for the Degree of

Bachelor of Technology

(Electronics and Communication)

: Presented & Submitted By :

PONNA DINESH

Roll No. (U20EC009)

B. TECH. IV(EC), 8th Semester

: Guided By :

Dr. Suman Deb

Assistant Professor, DoECE

(Year: 2023-24)

DEPARTMENT OF ELECTRONICS ENGINEERING

SARDAR VALLABHBHAI NATIONAL INSTITUTE OF TECHNOLOGY

Surat-395007, Gujarat, INDIA.

Sardar Vallabhbhai National Institute Of Technology

Surat - 395 007, Gujarat, India

DEPARTMENT OF ELECTRONICS ENGINEERING

CERTIFICATE

This is to certify that the Internship Training Report entitled "**Speech Emotion Recognition with Combination of CNN-LSTM and DNN**" is presented & submitted by PONNA DINESH, bearing Roll No. U20EC009, of B.Tech. IV, 8th Semester in the partial fulfillment of the requirement for the award of B.Tech. Degree in Electronics & Communication Engineering for academic year 2023-24.

He has successfully and satisfactorily completed **Internship Training Exam** in all respects. We certify that the work is comprehensive, complete, and fit for evaluation.

Dr. Suman Deb
Assistant Professor& Guide

Examiners:

Name of Examiners	**Signature with Date**
1. Dr. Jigisha N. Patel	____________________
2. Dr. Raghavendra Pal	____________________

Dr. J. N. Sarvaiya
Head, DoECE, SVNIT

Seal of The Department
(May 2024)

Internship Training Summary

(From 1st January 2024 to 26th April 2024)

Roll No: U20EC009 **Name: PONNA DINESH**

Sr No	**Joining Date**	**Relieving Date**	**Internship Organization**
1	1st January 2024	26th April 2024	SVNIT

Signature of the Internal Guide
UG Internship Training 2024

Acknowledgements

I would like to express my profound gratitude and deep regards to my guide Dr. Suman Deb for his guidance. I am heartily thankful for suggestions and the clarity of the concepts of the topic that helped me a lot for this work. I would also like to thank Prof.Dr. Jignesh N. Sarvaiya, Head of the Electronics Engineering Department, SVNIT and all the faculties of DoECE for their co-operation and suggestions. I am very much grateful to all my classmates for their support.

PONNA DINESH(U20EC009)
Sardar Vallabhbhai National Institute of Technology
Surat

May 2024

Abstract

Emotion is essential to all living things. Understanding emotion is challenging for everyone, but it can solve thousands of problems and save many lives if done correctly. Emotion is represented not only in gestures but also in work and the ability to produce effective results. As a result, throughout the past three decades, numerous researchers have been interested in emotion recognition through speech. In this study, addressed the crucial task of speech emotion recognition (SER) by leveraging deep learning architectures, namely Convolutional Neural Networks (CNN) combined with Long Short-Term Memory (LSTM) layers and Deep Neural Networks (DNN). My approach utilises a diverse feature set, including Mel spectrogram, MFCCs, spectrogram, delta of MFCCs, chroma feature, and tonal centroid features extracted from the EmoDB dataset. Through extensive experimentation employing a 5-fold cross-validation methodology, achieved notable classification accuracies of 77.38% with the CNN+LSTM model and 87.09% with the DNN model.

Additionally, investigated the impact of different hyperparameters and training strategies on the performance of the CNN+LSTM and DNN models, striving to optimise their accuracy and robustness in real-world applications. The experimental results revealed insights into the optimal configuration of these models, shedding light on potential avenues for further improvement. Moreover, explored the generalisation capability of the trained models by evaluating their performance on unseen data from diverse emotional speech datasets, thus ensuring the reliability and applicability of the proposed approach across different contexts. This comprehensive analysis enhances our understanding of SER methodologies and provides valuable guidance for future research endeavours to push the boundaries of emotion recognition technology.

Table of Contents

Page

Chapters

List of Figures

List of Tables

List of Abbreviations

1D CNN	1-Dimensional Convolutional Neural Network
2D CNN	2-Dimensional Convolutional Neural Network
3D CNN	3-Dimensional Convolutional Neural Network
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolution Neural Networks
CSV	Comma-Separated Values
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DNN	Deep Neural Network
ELU	Exponential Linear Unit
ELM	Extreme Learning Machines
FFT	Fast Fourier Transform
HCI	Human-computer interaction
KCV	K–Fold cross validation
LPC	Linear Prediction Coefficients
LSTM	Long Short-term memory
LIF	Local invariant features
LPCC	Linear prediction cepstrum coefficient
MFCC	Mel-frequency cepstral coefficients
MLP	Multi-Layer Perception
RDBN	Random Deep Belief Network
ReLU	Rectified linear unit
RMS	Root Mean Square
RNN	Recurrent Neural Network
SAE	Sparse auto-encoder
SDFA	Salient discriminative feature analysis
SER	Speech Emotion Recognition
SVM	Support Vector Machine
STFT	Short-Time Fourier Transform
TanH	Hyperbolic Tangent

Chapter 1
Introduction

Speech is one of the most fundamental and intuitive forms of human communication, not only conveying information but also expressing a wide array of emotions. Emotions play a pivotal role in interpersonal interactions, influencing behavior, decision-making, and social relationships. Recognizing these emotions accurately is therefore crucial for enhancing human-computer interaction (HCI). Speech Emotion Recognition (SER) seeks to automate the identification of emotional states from spoken language, using various acoustic and linguistic cues inherent in the human voice [7]. With the advent of sophisticated algorithms and computational models, SER has gained significant attention, promising transformative impacts across various sectors. In customer service, for instance, SER can be utilized to detect dissatisfaction or stress in voices, allowing for more responsive and tailored service. In healthcare, monitoring emotional cues in speech can assist in diagnosing and treating conditions such as depression and anxiety. Moreover, in educational technologies, understanding the emotional state of learners can help in adapting content delivery to improve engagement and learning outcomes.

Emotions can be gathered and appraised in various ways using technology, including facial expressions, physiological cues, and voice, all examples of nonverbal communication. Identifying and appropriately handling emotions conveyed through signals is necessary to enable more intuitive and natural communication between humans and computers. Over the last 20 years, much research on automatic emotion recognition has led to developing and improving numerous machine learning algorithms. Emotion recognition is applied in numerous applications. Anger detection can assess the quality of voice portals or contact centres. It enables service providers to tailor their offerings to the emotional states of their customers. Monitoring the stress of aircraft pilots can assist in lowering the risk of an aircraft accident in civil aviation. Many researchers have incorporated the emotion detection module into their products to improve players' experiences with video games and keep them motivated.

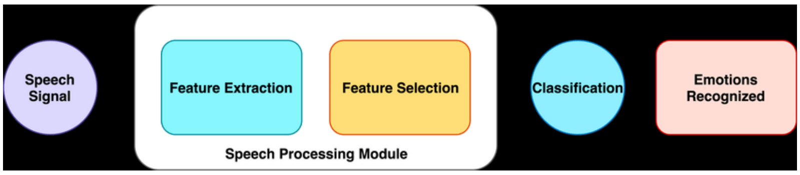


Figure 1.1: Traditional Speech Emotion Recognition System [1]

1.1 History

The history of speech emotion recognition (SER) using Deep Neural Networks (DNN) dates back to the early 2000s, with pioneering works such as those by Deng et al. [8] and Eyben et al. [9], who demonstrated the effectiveness of DNN architectures in extracting discriminative features from speech signals for emotion classification. Concurrently, the combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks emerged as a powerful approach for SER, as evidenced by the work of Han et al. and Weninger et al. [10]. These studies leveraged the hierarchical feature learning capabilities of CNNs for extracting spectral and temporal features from speech spectrograms, followed by LSTM networks to capture long-range dependencies and temporal dynamics in the emotional content of speech. Since then, the fusion of CNNs and LSTMs has become a prevalent paradigm in SER, offering enhanced performance in emotion recognition tasks by effectively modelling local and global speech data dependencies.

Many scholars presented a range of strategies in this domain in SER, a new research area. Most researchers focus on identifying influential, prominent, and discriminative characteristics of speech signals for categorisation to detect a speaker's emotion accurately. Researchers have recently applied deep learning algorithms to determine SER's salient and discriminative characteristics. Convolutional neural networks build high-level features on top of low-level characteristics to perceive and distinguish lines, dots, curves, and forms. Compared to low-level handcrafted features, deep learning models (CNN, CNN-LSTM, DNN, DBN, and others) methods to detect high-level salient features gain superior accuracy. The ultimate goal is to provide insights into the efficacy of deep learning techniques for speech emotion recognition, with the aim of advancing applications in human-computer interaction, emotional analysis, and assistive technologies.

1.2 Problem Statement

This project aims to develop a precise system for Speech Emotion Recognition utilizing a combination of CNN+LSTM and DNN models, with a focus on leveraging the EmoDB dataset and employing a 5-fold cross-validation approach. The objective is to evaluate the performance of these models in accurately detecting and categorizing emotional states conveyed through speech signals. Key challenges include optimizing model architecture for effective feature extraction, enhancing classification accuracy across diverse emotional categories, and ensuring robustness to variations in speech patterns and environmental conditions.

1.3 Objective

Developing a precise and accurate system for speech emotion recognition using CNN+LSTM models, aiming to overcome various challenges associated with effectively capturing and classifying emotional cues from speech signals. This project focuses on addressing the following challenges:

1. **Model Development:** Develop CNN+LSTM and DNN models for speech emotion recognition, leveraging their complementary strengths in capturing temporal and spatial features within speech signals.

2. **Feature Extraction:** Extract relevant features from speech signals in the EmoDB dataset, including Mel spectrograms, MFCCs, Delta and Delta-Delta MFCCs, Spectrogram, Chroma and Tonnetz, to enable effective emotion classification.

3. **Model Evaluation:** Utilize 5-fold cross-validation to rigorously evaluate the performance of the CNN+LSTM and DNN models, ensuring robustness and generalization across different folds of the dataset.

4. **Comparative Analysis:** Compare the effectiveness of the CNN+LSTM and DNN models in accurately recognizing and classifying emotional cues from speech signals, identifying the model that yields the highest performance metrics.

5. **Emotional Classifications:** Train the models to classify a broad spectrum of emotions present in the EmoDB dataset, including but not limited to happiness, sadness, anger, fear, and neutrality.

6. **Optimization:** Explore techniques for optimizing model architecture, hyperparameters, and feature representation to enhance classification accuracy and computational efficiency.

1.4 Motivation

Improving human-computer interaction (HCI) through technology, skill development, deep learning, and artificial intelligence techniques is necessary to improve HCI, including emotion recognition. SER is a valuable area of HCI with various real-time applications, including virtual reality, assessing user happiness in contact centres, identifying human emotion in human reboot interactions, and assessing a user's emotional state to determine the proper response in emergency call centres. The SER for the automobile board system is designed to identify drivers' emotional or mental states so that necessary safety measures can be taken for the passengers.

Moreover, SER is utilised in automatic translation systems and detecting violent and destructive behaviour in crowds that cannot be completed by hand [11]. SER for intelligent, efficient services to offer methods for showcasing the efficiency of the SER for intelligent healthcare facilities using CNN architectures with rectangular shape filters [12]. The effectiveness of SER for real-time applications that use discriminative and salient features analysis to extract features to increase the significance of HCI.

1.5 Organization of the Report

The report is structured into five chapters, starting with an introduction that provides an overview of speech emotion recognition. The first chapter lays the foundation for understanding the significance and context of the study. The subsequent chapter, the Literature Review, delves into existing research, frameworks, and methodologies related to speech emotion recognition, highlighting key findings and gaps in knowledge. Following this, the Methodology chapter outlines the approach and techniques employed in the study, elucidating the data collection, preprocessing, feature extraction, and the chosen model for analysis. The fourth chapter presents the Results and Analysis section, where the study outcomes are presented, analysed, and discussed in detail. Finally, the Conclusion and Future Scope chapter summarises the findings, reiterates the significance of the study, and provides insights into potential future research directions and advancements in speech emotion recognition.

Chapter 2
Literature review

Enhancing the communication between humans and machines requires significant work in the area of emotion recognition. Emotional complexity increases the difficulty of the acquisition task. Below are some literature reviews on various technologies and numerous approaches, along with citations.
The study conducted by Siba Prasad Mishra et al. delves into the challenging yet crucial domain of emotion recognition through speech signals. Siba Prasad Mishra et al. contribute to this field by employing a combination of advanced techniques, including mel frequency cepstral coefficients (MFCC), spectrogram, and mel-spectrogram, as features for classification. Their study utilizes convolutional neural networks (CNN) and deep neural networks (DNN) to analyze these features individually and in combination. Notably, the fusion of features using both CNN and DNN classifiers significantly enhances emotion classification accuracy compared to using individual features alone. Moreover, their proposed feature and classifier-level fusion method outperform existing state-of-the-art approaches, showcasing the efficacy of their methodology in accurately recognizing speech emotions across various datasets. Through their comprehensive literature review and empirical findings, Siba Prasad Mishra et al. underscore the importance of SER and present a promising avenue for further advancements in this critical domain [2].

Dario Bertero et al. described their approach of enabling an interactive dialogue system to recognize user emotion and sentiment in real-time. These modules enable ordinarily standard dialogue systems to show “empathy” and respond to the user while being aware of their emotion and intent. Emotion identification from speech was formerly achieved by feature engineering and machine learning, with the first stage causing a delay in decoding time. They described a CNN model that extracts emotion from raw speech input without using feature engineering. This approach outperforms traditional feature-based SVM classification, with an average accuracy of 65.7% across six emotion categories, a 4.5% increase. A separate, CNN-based sentiment analysis module distinguishes sentiments from speech recognition results, with an F-measure of 82.5 on human-machine dialogues when trained using out-of-domain data [13].

Ashima Yadav et al. proposed A Multilingual Framework of CNN and Bi-LSTM for Emotion Classification. Suggested a language-independent deep learning framework for voice emotion classification. Developed a novel mix of 1D CNN and Bi-LSTM units to extract both MFCC-based and deep high-level features. The proposed system uses CNN to extract local information from signals, while the Bi-LSTM layer models the signal's long-term contextual dependencies. They verify their suggested architecture against two multilingual datasets, EmoDB for German and RAVDESS for English [14].

Guihua Wen et al. [15] proposed Random Deep Belief Networks for Recognizing Emotions from Speech Data, which describes an ensemble learning technique for recognising emotions from speech signals using Random Deep Belief Networks (RDBN). After collecting the low-level features of the input speech stream, they applied the Random subspaces method. Where each Random subspace is sent into the DBN input, which extracts the higher-level properties of the input voice signal and feeds these higher-level features to the base classifier, which generates a projected emotion label. Furthermore, each outputted emotion label is fused with the final emotion label for the given input speech signal via majority voting.

M. Shamim Hossain and Ghulam Muhammad [16] proposed an emotion identification system based on deep learning using Audio-Visual emotional big data that shows how emotions may be recognised utilising speech and video as input. For this, they used two datasets: a Big Data database containing both audio and video input files, and the eNTERFACE database. In this technique, they first retrieved the properties of the given input speech signals to build a Mel-spectrogram, which can be thought of as a picture. This Mel-spectrogram is transmitted to a 2D CNN followed by extreme learning machines for score fusion (ELMs). When it comes to video signals, specific sample frames from a video segment are gathered and fed into a 3D CNN, which is subsequently followed by extreme learning machines (ELMs) for score fusion. The output of both these speech and video fusions is fed into an SVM for final emotion categorization of the input speech and video signals.

Mingke Xu et al. [17]created a speech emotion identification system based on multiscale area attention and data augmentation in which they employed multiscale area attention to improve speech emotion detection and built an attention-based CNN. They began by extracting features from the log Mel- spectrogram using the Librosa package. These features were then fed into two concurrent convolutional layers to produce textures for the time and frequency axes respectively. The output is sent through four convolutional layers in a row, resulting in an 80-channel representation. After that the attention layer pays attention to the representation and sends the results to the fully connected layer for final emotion categorization. In addition, SVM and its combination techniques have been applied in several studies. However, in the context of ensemble learning the current state of the art for speech emotion recognition does not combine random subspace , CNN, MLP.

Mujaddidurrahman et al. [18] proposed this literature on SER. Speech emotion recognition has become a difficult problem, particularly in human-machine interaction. Each person communicated their feelings in different ways, and the characteristics of speech are still unclear in order to differentiate between existent emotions.Speech results were created by mental and psychological states, which are directly influenced by emotions.This study presented audio emotion recognition using a 2D-CNN model with a Log-Mel spectrogram as input.The suggested 2D-CNN model with Log-Mel spectrogram technique successfully captured substantial voice data.The EMODB database was used in the trials to evaluate the proposed model with supplemented data.The suggested model outperformed current models, which also used deep learning approaches to recognize emotions in speech in terms of accuracy.

Fushiki [19] proposed Estimation of prediction error by using K-fold cross-validation. When making predictions is our goal, estimating prediction accuracy is crucial. Although the training error has a downward bias, it is a simple way to estimate prediction error. K-fold cross-validation, on the other hand, has an upward bias.In leave-one-out cross validation, the upward bias might not be significant, but in 5-fold or 10-fold cross validation—which is preferred from a computational perspective—it occasionally cannot be disregarded.There will be an appropriate estimate in a family that connects the two estimates since K-fold cross-validation has an upward bias and the training error has a downward bias.In this paper, they looked into two families that have a relationship between K-fold cross-validation and training error.

Aida-zade, Xocayev and Rustamov [20]proposed on Speech recognition using Support Vector Machines. For an Azerbaijani data set, they employed Support Vector Machines to build an acoustic model of a Speech Recognition System based on MFCC and LPC features. A Multilayer Artificial Neural Network that used this DataSet to recognize speech produced some results. The application of SVM techniques to the Azerbaijan Speech Recognition System is the primary objective of this work. The range of SVM results with various Kernel functions is examined during the training phase. It is shown that SVM with polynomial kernels and radial basis performs better in recognition than Multilayer Artificial Neural Network.

Anguita, Davide and Ghio, Alessandro and Ridella, Sandro and Sterpi, Dario reviewed the k–Fold cross validation (KCV) technique,applied to the Support Vector Machine(SVM) classification algorithm. They compared a number of KCV technique variations, some of which are more rigorous in identifying the correct classifier but are less frequently used by practitioners. Some of the variations lack theoretical support.The latter ones enable the establishment of an upper bound on the SVM's error rate, which is a means of statistically ensuring the classifier's dependability and is, thus, highly significant in numerous real-world applications [21].

Emotions are an integral part of human interactions and are significant factors in determining user satisfaction or customer opinion. Modules for speech emotion recognition (SER) are also crucial to the advancement of applications for human–computer interaction (HCI). Over the past few decades, an enormous number of SER systems have been developed. Deep neural networks (DNNs) that are attention-based have proven to be effective tools for extracting information from multimedia content that is time-dispersed unevenly. In order to emphasize emotionally salient information, DNN architectures have recently included the attention mechanism. In addition to reviewing recent advances in SER, this paper looks at how different attention mechanisms affect SER performance. An extensive analysis of the system accuracies is conducted using the popular IEMOCAP benchmark database [1].

Q. Mao, M. Dong, Z. Huang, and Y. Zhan [12] proposed Learning salient features for speech emotion recognition using convolutional neural networks. The use of convolutional neural networks (CNN) to acquire affect-salient features for SER was suggested in this paper. Simple features are learned in the lower layers of CNN, whereas affect-salient, discriminative features are obtained in the higher layers. There are two separate learning phases in CNN. To learn local invariant features (LIF) from unlabeled samples, a version of the sparse auto-encoder (SAE) with reconstruction penalisation is employed in the first step. In the second step, the local invariant features are fed into a feature extractor called salient discriminative feature analysis (SDFA), which identifies affect-salient, discriminative features. They proposed a novel objective function in SDFA by encouraging feature saliency, orthogonality, and discrimination for SER.

Emotion recognition from speech signals in human-machine interface applications has long been a research area. Many systems have been developed to identify the emotions from the speech signal. This paper reviews speech emotion recognition based on earlier technologies that employ various classifiers for emotion recognition. Classifiers distinguish between emotions, including neutral, surprise, happiness, sadness, and anger. Emotional speech samples serve as the database for the speech emotion recognition system. Energy, pitch, linear prediction cepstrum coefficient (LPCC), and Mel frequency cepstrum coefficient (MFCC) are features extracted from these speech samples. Extracted features serve as the basis for the classification performance. Also covered are conclusions regarding the functionality and constraints of the speech emotion recognition system based on various classifiers [22].

Khalil, Ruhul Amin and Jones, Edward and Babar, Mohammad Inayatullah and Jan, Tariqullah and Zafar, Mohammad Haseeb and Alhussain, Thamer [23] proposed Speech Emotion Recognition Using Deep Learning Techniques. Emotion recognition from speech signals is an essential but challenging human-computer interaction (HCI) feature. The literature on speech emotion recognition (SER) has employed a variety of approaches, including well-known speech analysis and classification techniques, to extract emotions from signals. Deep learning techniques have recently been proposed as an alternative to traditional machine learning techniques. This paper presents an overview of deep learning techniques for speech-based emotion recognition and discusses recent research that uses them. The review discusses the extracted emotions, the databases used, the improvements made to speech emotion recognition, and its shortcomings.

2.1 Research Gap

From this Literature review, I got to know that there are several techniques, mainly DNN with an accuracy of 87%, SVM with an accuracy of 44.4% with the Berlin database, CNN with an accuracy of 75.94% with the Emo-DB dataset, 82.31% for eight emotions and 79.42% for five emotions on the IEMOCAP and Emo-DB datasets, respectively, using a CNN-Transformer architecture for capturing spatial and sequential features,82% on the RAVDESS dataset using a combined CNN-LSTM architecture and 74% using a CNN-Transformer encoder architecture. The literature review shows that advanced deep learning techniques, such as CNNs, RNNs, and their hybrids, have shown significant promise in improving speech emotion recognition (SER) accuracy compared to traditional methods like SVM. However, despite the advancements achieved, several research gaps warrant further investigation. Firstly, while some studies report high accuracies on specific datasets, such as the Emo-DB or IEMOCAP, there needs to be more consistency across different databases and languages. This suggests a need for more robust and generalised models that can effectively handle variations in speech signals across diverse contexts.

Chapter 3
Proposed Methodology

In this chapter, the methodology of Speech Emotion Recognition is discussed. Firstly, speech processing is addressed; for this speech processing, a dataset of different audio signals recorded by professional speakers was used [24]. It begins by importing necessary libraries and defining functions for energy and RMS calculations. It then loads the EmoDB dataset, a collection of German emotional speech, and normalises the audio signals. Several features, including Mel spectrogram, MFCCs, spectrogram, the delta of MFCCs, chroma feature, tonal centroid features (tonnetz), spectral contrast, and RMS energy, are extracted from these signals.

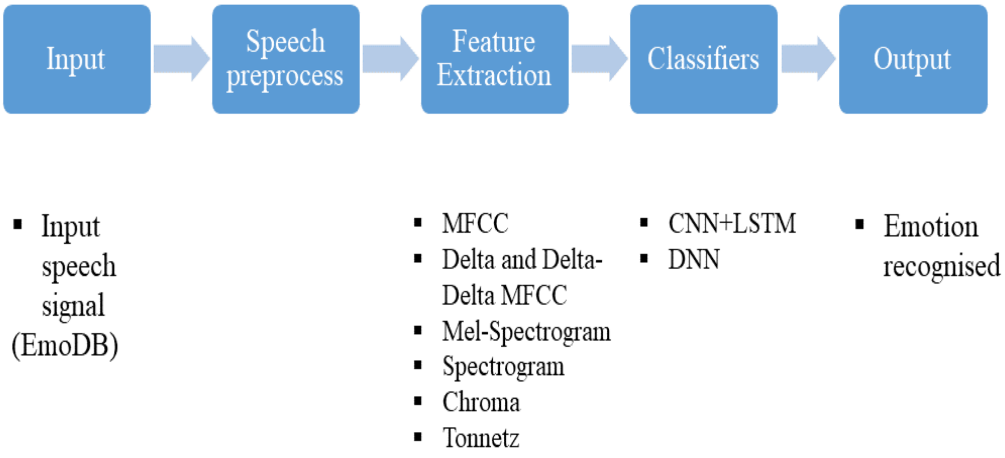


Figure 3.1: Block diagram of Speech Emotion Recognition

These features, commonly used in speech and audio processing, are appended to a list along with the filename and emotion label. Finally, this data is written into a CSV file, which can be used as input for machine learning models for emotion classification. The effectiveness of these features can vary depending on the task and dataset, so experimentation and normalisation are recommended. A Deep Neural Network (DNN) model based on the EmoDB dataset was proposed to classify emotions. The dataset is divided into training and testing sets, and the unique values of the target variable are identified. The class weights are computed to handle class imbalance. The features are then standardised using the StandardScaler.

The DNN model has four layers, each followed by batch normalisation and dropout for regularisation. The model is compiled with the Adam optimiser, sparse categorical cross entropy as the loss function, and sparse categorical accuracy as the metric.

The model's initial weights are saved and then returned to ensure reproducibility. The model is trained for 1200 epochs with a batch size 64, using early stopping and model checkpoint callbacks. The best model is saved based on the validation sparse categorical accuracy. It uses a Convolutional Neural Network (CNN) combined with Long Short-Term Memory (LSTM) layers to classify emotions based on the EmoDB dataset. The dataset is divided into training and testing sets, and the unique values of the target variable are identified. The class weights are computed to handle class imbalance. The features are then standardised using the StandardScaler. The model is built with several layers: four Conv1D layers, each followed by MaxPooling1D, BatchNormalization, and Dropout for regularisation, two LSTM layers, and three Dense layers. The model is compiled with the Adam optimiser, sparse categorical cross entropy as the loss function, and sparse categorical accuracy as the metric. The model's initial weights are saved and then returned to ensure reproducibility. The model is trained for 900 epochs with a batch size 64, using early stopping and model checkpoint callbacks. The best model is saved based on the validation sparse categorical accuracy. In the current project used the 5-fold cross-validation strategy.

3.1 Database

Emotional Speech Databases: Emotional speech databases play a crucial role in advancing speech emotion recognition systems. These databases contain human speech recordings specifically designed to capture emotional content. Here are the three main types of emotional speech databases:

3.1.1 Natural Databases

Natural databases contain spontaneous speech samples recorded in real-world settings, capturing genuine emotional expressions. Researchers often collect these recordings without participants' awareness, ensuring authenticity [25].

3.1.2 Simulated (Acted) Databases

Simulated databases involve controlled scenarios where actors intentionally express specific emotions. These actors follow predefined scripts or instructions to simulate emotional states. Such databases allow researchers to systematically study emotions and create consistent training data [26].

3.1.3 Elicited (Induced) Databases

Elicited databases are created by eliciting emotions from participants through specific stimuli or tasks. Researchers induce emotions using visual cues, storytelling, or other techniques. These databases provide valuable insights into how external factors influence speech patterns related to emotions [27].

3.2 Dataset

The EmoDB dataset, also known as the Berlin Database of Emotional Speech, is a freely available German emotional database1. It was created by the Institute of Communication Science at the Technical University in Berlin, Germany1. The dataset comprises 535 utterances recorded by ten professional speakers, five males and five females. The EmoDB dataset encompasses seven emotions: anger, boredom, anxiety, happiness, sadness, disgust, and neutral [28]. The data was initially recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz1 [28]. This dataset is commonly used for classification problems related to emotional speech [29] and has been leveraged in various research studies for tasks such as speech emotion recognition [28]. Each utterance is given a name based on the same system. For instance, the audio file 03a01Fa.wav contains speaker 03's speech of text a01 with the emotion "Freude" (Happiness).

3.3 Speech Features for Emotion Recognition

3.3.1 Mel-frequency cepstral coefficients - MFCC

The calculation of Mel-Frequency Cepstral Coefficients (MFCC) is viewed as an artificial method aimed at mimicking human auditory perception, particularly in identifying speech and emotional states. It utilises frequency filters placed in a linear configuration at lower frequencies and logarithmically at higher frequencies. Such a configuration is essential for preserving the critical phonetic components of speech signals, which are vital for creating MFCC attributes. Tone frequencies in voice transmissions are measured in Hz, but the Mel scale determines subjective pitch. Speech signals frequently contain tones of varying frequency. The Mel-frequency scale uses linear spacing below 1000 Hz and logarithmic spacing beyond this frequency. Equation (3.1) provides the formula for converting frequency to Mel-frequency, with Mel(f) representing the Mel frequency and f being the frequency in Hz.

$$Mel(f) = 2595 * log(1 + \frac{f}{700}) \tag{3.1}$$

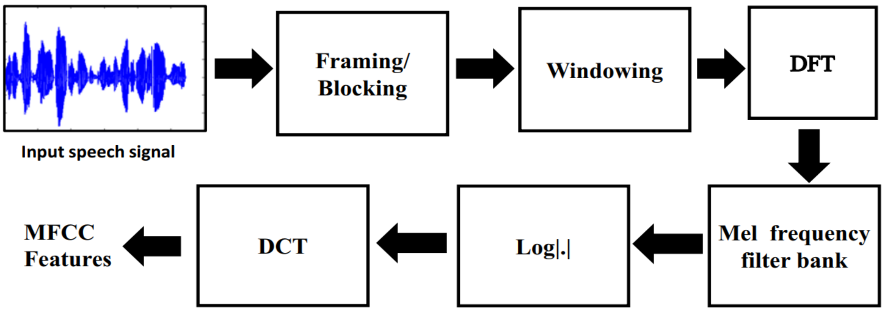


Figure 3.2: Mel-frequency cepstrum coefficients (MFCCs) block diagram [2]

In this case, f is the frequency (Hz), and Mel(f) is the frequency (mels). Figure 3.2 shows the block diagram utilised for MFCC feature extraction. A voice signal is non-stationary because its structure changes with time, although it behaves quasi-stationary for a short period. As a result, speech analysis must be conducted on a brief slice of speech. To extract MFCC features, the speech signal was first put through a pre-emphasis filter, which balanced the amplitude spectrum of spoken sounds in the upper-frequency range. The transfer function for the pre-emphasis filter is given by Equation (3.2).

$$H(z) = 1 - bz^{(-1)} \tag{3.2}$$

Where b is the emphasised parameter between 0.9 and 1, each audio frame is subjected to an overlapping window using either a Hanning or Hamming technique. This step enhances the harmonic components and smoothens the boundary transitions to minimise discontinuities and optimise spectral resolution. The next step involves applying the discrete Fourier transform (DFT) to the signal. The frequency analysis is then adapted to the mel scale using a mel-filterbank to align with human auditory perception. This adaptation is crucial because it reflects human listeners' logarithmic perception of pitch and frequency. Following this, a discrete cosine transform (DCT) is performed to reduce spectral leakage and highlight the spectrum's amplitude characteristics. The final output is the Mel Frequency Cepstral Coefficients (MFCC), effectively capturing the sound's timbral aspects [2].

3.3.2 Delta and Delta-Delta MFCCs

Delta and Delta-Delta MFCCs, derived as the first and second derivatives of Mel Frequency Cepstral Coefficients (MFCCs), capture temporal dynamics within speech signals. While MFCCs represent spectral features at a single time frame, the deltas provide information about the rate of change akin to velocity. The delta-deltas, or acceleration coefficients, reveal how quickly these features evolve. By incorporating these derivatives, create an extended feature vector characterising static spectral properties and tracing their trajectory over time. This comprehensive approach enhances the robustness and accuracy of speech recognition systems, particularly in noisy and dynamic real-world scenarios.

3.3.3 Mel-Spectrogram

Human auditory perception typically regards pitches on the Mel scale as equidistant. As the frequency, measured in Hertz, increases, so does the interval between values on the Mel scale. The mel-spectrogram is a tool that converts Hertz values into the mel scale, aligning with how humans perceive sound. While linear audio spectrograms are optimal when all frequencies need equal representation, mel-spectrograms are better suited for mimicking human auditory perception, displaying frequencies beyond a specific threshold on a logarithmic scale. In a mel-spectrogram, brighter red hues indicate stronger amplitudes, while darker blue hues suggest lower amplitudes or quieter parts of the speech signal. Mel-spectrograms are invaluable in various speech-related applications, including sound classification, speaker identification, speech pathology, and emotion detection [2].

3.3.4 Spectrogram (STFT)

The discrete Fourier transform (DFT) is applied to audio signals frame-by-frame, translating them from the time domain to the frequency domain. This transformation details the frequencies present in the signal but loses the temporal information. To overcome this limitation, a spectrogram is used, which captures the audio signal's time and frequency information. A spectrogram visually maps the intensity of the signal at various frequencies over time within a single waveform. This representation allows one to observe how the signal's energy fluctuates across different frequencies and how these changes evolve throughout the recording. It is typically displayed on a two-dimensional graph where time is plotted on the horizontal axis, frequency on the vertical axis, and the strength or loudness of various frequency components is indicated through colour or intensity at each point [2].

3.3.5 Chroma

Chroma features, chroma vectors or chromatograms, are essential in information retrieval, focusing on music's harmonic and melodic aspects. These features simplify an audio signal's spectrum into 12 pitch classes corresponding to the musical octave, thus capturing harmony, chords, and tonal elements. The process involves transforming the audio to the frequency domain using the Fast Fourier Transform (FFT), then mapping spectral data to these pitch classes by summing contributions across octaves, facilitating tasks like chord recognition and music genre classification [30]. Chroma features are robust against timbral and dynamic changes, making them suitable for analysing polyphonic music where multiple instruments are present, supporting applications in automatic music transcription and recommendation systems [31].

3.3.6 Tonnetz

Tonnetz features leverage the concept of the “Tonnetz” (German for “tone network”), a lattice diagram used to explore tonal space and harmonic relationships within music theory. These features map pitch classes onto a two-dimensional geometric space where notes are positioned based on their harmonic closeness, such as fifths and thirds, aiding tasks like harmonic analysis and key detection [32]. In computational music analysis, Tonnetz highlights the geometric structure of music, enabling the study of chord progressions and tonal transformations, which are functional for applications in music education and automated music generation. The software can predict harmonic changes and recognise patterns by employing algorithms that interpret the Tonnetz model, making theoretical music concepts visually and audibly accessible [33].

3.4 Feature Extraction

Now, the audio sample’s feature is extracted using the Librosa library. ‘Librosa’ is the Python package for audio analysis and music analysis. It provides building blocks that are necessary to create information of music of retrieval systems. It is a Python package for music analysis. It is used in Python to extract and process features from audio files. It starts with loading each audio file and normalising the signal. This normalised signal extracts a variety of features that are commonly used in audio processing. These include the Mel spectrogram, MFCCs, and signal spectrogram. It also calculates the delta and delta-delta MFCCs, which capture the change in MFCCs over time. Other features extracted include chroma (relating to the 12 different pitch classes), tonnes (harmonic relations between different pitches), spectral contrast (difference in amplitude between peaks and valleys in a sound spectrum), and RMS energy. These features are then averaged over time to create a single feature vector for each audio file. This feature vector and the label for the emotion present in the audio file are then written to a CSV file. This results in a preprocessed dataset where each instance corresponds to an audio file and consists of the extracted features and the emotion label.

3.5 Cross-Validation

Cross-validation is a valuable method for assessing a statistical model's efficiency. A training and validation set are the two subsets created by splitting the available data into them. The validation set is used to evaluate the model's performance after it has been trained on the training set. This procedure guarantees a comprehensive assessment of the model's capacity to generalise to unknown data. Through iterative training and evaluation of various subsets of the data, cross-validation yields a more reliable estimate of the model's actual performance. This method aids in preventing overfitting, a phenomenon in which a model becomes unduly adapted to the training set and is unable to make an excellent generalisation to fresh data.
The typical techniques for cross-validation are :

1. Validation Set Methodology
2. Cross-validation with leave-P-out
3. Do not include one cross-validation
4. Cross-validation in K-fold
5. K-fold stratified cross-validation

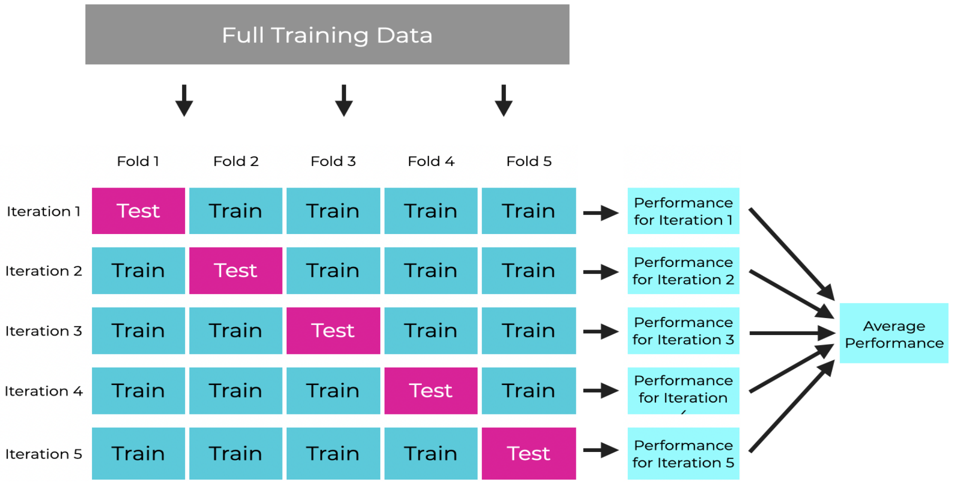
3.5.1 K-fold cross-validation

One method that is frequently used to assess how well prediction models work is K-fold cross-validation. The training data is divided into k folds or equal-sized chunks. The model is trained on k-1 folds for each iteration, and its performance is assessed on the remaining fold. To ensure that every fold functions as training and testing data, this step is performed k times. When analysing the model's generalizability, the average performance over all folds offers a more thorough evaluation than testing it on a single holdout set.
The steps for k-fold cross-validation are:

- The input dataset is created into k groups
- Among the k groups, one group is selected as the test dataset
- Rest of the groups are considered as the training dataset
- Now the model is fit with the training dataset, and the performance of the model is evaluated on the test dataset

For example, consider the 5-fold cross-validation:

- Five folds are used to divide the dataset. The first fold in the first iteration is for the test data, and the remaining folds are for training. The test data for the second iteration comes from the second fold, and the remaining folds are used for training. All of the folds will go through this procedure once more. The 5-fold cross-validation diagram is shown in Figure.3.3.



	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	
Iteration 1	Test	Train	Train	Train	Train	Performance for Iteration 1
Iteration 2	Train	Test	Train	Train	Train	Performance for Iteration 2
Iteration 3	Train	Train	Test	Train	Train	Performance for Iteration 3
Iteration 4	Train	Train	Train	Test	Train	Performance for Iteration
Iteration 5	Train	Train	Train	Train	Test	Performance for Iteration 5

Figure 3.3: 5-fold cross-validation diagram [3]

3.6 Classifiers

A classification system assigns a specific emotion class to each speech based on the retrieved elements from the speech. Various classifiers are available for emotion recognition.

3.6.1 Deep Neural Network - DNN

A primary neural network comprises three layers: the input layer, one hidden layer, and the output layer. However, a Deep Neural Network (DNN) contains multiple hidden layers. The input layer receives data from external sources and forwards it to the first hidden layer. Each hidden layer comprises neurons connected by weights to the preceding and subsequent layers. The primary function of a neuron is to compute a weighted sum of the inputs it receives and process this sum through an activation function. The result of the activation function is then passed on to the neurons in the following hidden layer. DNNs utilise various activation functions, such as TanH, ReLU (rectified linear unit), sigmoid, and ELU (exponential linear unit). The output layer, which has several neurons corresponding to the number of classes in the dataset, employs the softmax activation function to generate a predictive score for each class or emotional state in the voice signal.

Figure 3.8 depicts the block diagram of the proposed DNN model employed in my investigation, including parameter details. I used a DNN classifier with four hidden layers between the input and output layers. The first hidden layer contains 999 neurons, followed by batch normalisation and a dropout of 0.1. ELU is an activation function applied to all layers except the output layer. Equation (3.3) gives the ELU activation function.

$$f(x) = \begin{cases} a(e^{x}-1) & \text{for} \quad x < 0 \\ x & \text{for} \quad x \geq 0 \end{cases} \tag{3.3}$$

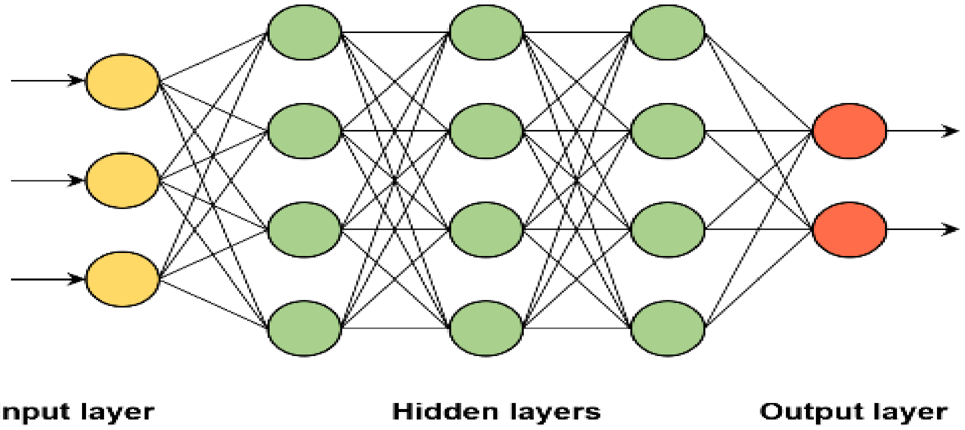


Figure 3.4: Simple architecture of deep neural network [4]

The output of the first hidden layer is passed into the second hidden layer, which has 785 neurons, followed by batch normalisation and a dropout of 0.2. The third and fourth hidden layers contain 721 and 672 neurons, respectively. Following batch normalisation, there is a dropout of 0.2 and 0.3. The output layer consists of seven neurons and predicts the score for each class using the softmax activation function. I applied the Adam optimiser with a learning rate 0.001 in this study.

3.6.2 Convolutional Neural Network - CNN

A Convolutional Neural Network (CNN) is a deep learning network that predominantly uses convolutional operations. Its architecture is divided into three principal layers: convolutional, fully connected, and output layers. During the convolutional stage, various filters are applied to the input data, executing convolution operations that result in outputs of high dimensionality, which vary with the filter count. To handle these dimensions, techniques like max pooling, min pooling, and average pooling are used to reduce the output size. After the convolutional stage, the processed data is transferred to a flattened layer, which reshapes the data into a one-dimensional feature vector. This vector becomes the input for the dense layers comprising numerous neurons that process and pass the information. The network's final dense layer, corresponding to the number of classes in the dataset, employs a softmax activation function to derive a probability distribution for each class, estimating the likelihood of each based on the processed inputs.

CNN Architecture

Multiple layers comprise a convolutional neural network, including the input, pooling, convolutional, and fully connected layers.

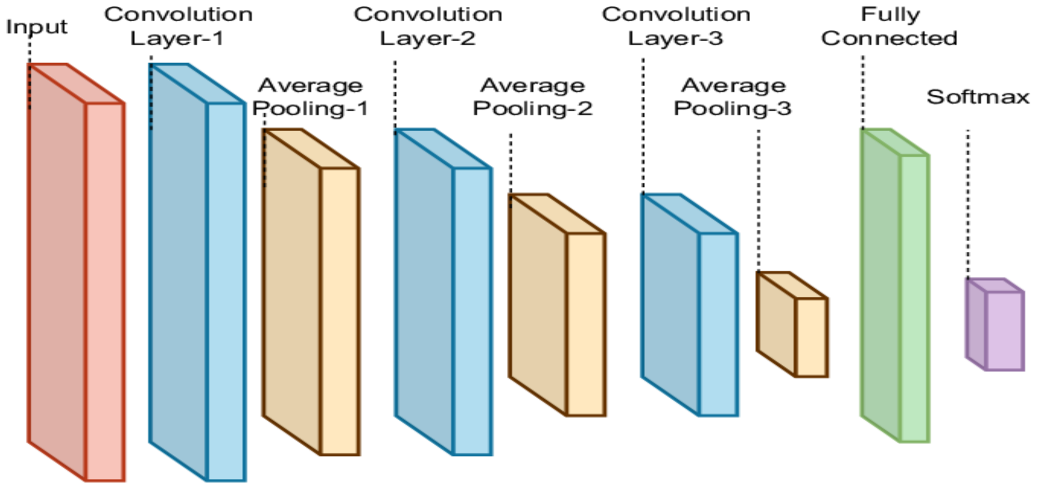


Figure 3.5: Basic architecture of the 1D-CNN [5]

- **Input Layers:**This layer is where we feed data into our model.

- **Pooling layer:** The pooling layer, which comes after the convolutional layer, decreases the dimensionality of each feature map while preserving crucial information. Various types of pooling can be employed, including Max, Average, and Sum pooling.

- **Convolutional Layer:** This layer performs a dot product of the input and a weight matrix, called a 'filter' or 'kernel', sliding over the input data (performing the convolution operation) and outputs feature maps.

- **Activation Layer:** Activation functions, on the other hand, determine the output of a neuron based on its inputs, introducing non-linear characteristics to the network.

- **Fully Connected Layer:** Following the pooling layer is the fully connected layer, which establishes connections between every neuron in one layer and every neuron in another. This layer functions similarly to the traditional multi-layer perceptron neural network (MLP).

3.6.3 Long Short-Term Memory - LSTM

Long Short-Term Memory (LSTM) networks are a form of recurrent neural network (RNN) that can learn order dependence in sequence prediction tasks. This is especially beneficial in jobs where the context provided by previous items in the sequence is critical to understanding or anticipating the following parts. LSTMs, invented in 1997 by Sepp Hochreiter and Jürgen Schmidhuber, were intended to address the limitations of typical RNNs, notably those related to learning long-term dependencies [34].

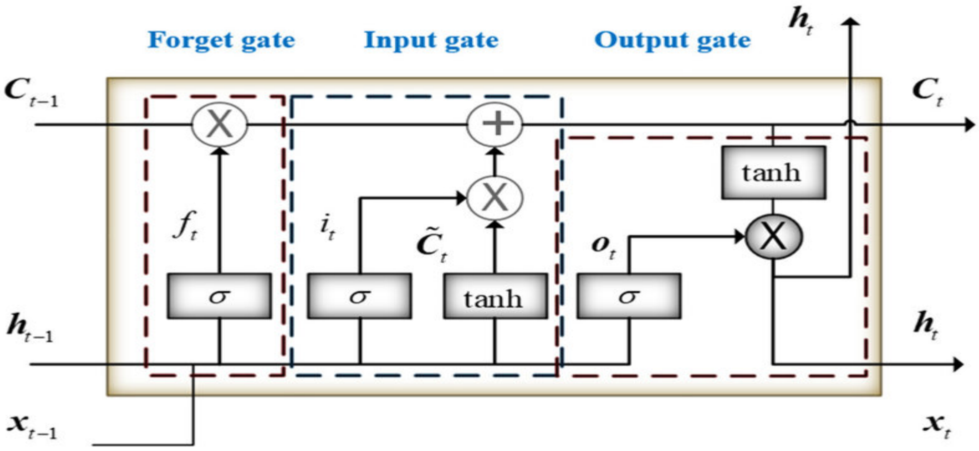


Figure 3.6: Basic-structure-of-a-long-short-term-memory-LSTM [6]

The structure of an LSTM is designed to avoid the long-term dependency problem typical of standard recurrent neural networks. At its core, it includes a cell state, which acts as a conduit for carrying information across sequence steps without alterations, and three types of gates: the input gate, the forget gate, and the output gate. These gates, essentially neural networks with sigmoid activation functions, regulate the information flow into and out of the cell state, deciding what to retain or discard, thus enabling the network to maintain or forget information dynamically over time. This architecture allows LSTMs to capture long-term dependencies within input sequences effectively.

3.7 Proposed CNN+LSTM And DNN Classifier

The study proposed a CNN+LSTM and DNN model to analyze emotion classification performance using the speech signal. The block diagram of the proposed CNN+LSTM and DNN classifier is shown in Figure.3.7 and Figure.3.8, respectively. The proposed combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architecture for emotion classification from speech signals comprises multiple layers tailored to extract and classify features effectively. Initially, the model employs convolutional layers with 256 filters, followed by max-pooling, batch normalization, and dropout for feature extraction. Subsequently, LSTM layers with 256 units capture temporal dependencies, and two dense layers with 256 neurons each further process the features. A dropout layer is applied after the first dense layer for regularization. Finally, the output layer with softmax activation predicts the probabilities of seven different emotions. This architecture integrates CNN for spatial feature extraction and LSTM for temporal feature learning, allowing the model to accurately classify emotions in speech data while mitigating overfitting through regularization techniques such as batch normalization and dropout.

The input to the proposed Deep Neural Network (DNN) model is a feature vector of size 260 extracted from the EmoDB dataset. The proposed DNN model includes multiple dense layers. The first dense layer contains 999 neurons and uses the Exponential Linear Unit (ELU) activation function, followed by a Batch Normalization layer and a Dropout layer with a rate of 0.1. The subsequent dense layers contain 785, 721, and 672 neurons, each followed by a Batch Normalization layer and a Dropout layer with rates of 0.2, 0.2, and 0.3, respectively. The output of the final Dropout layer is given to the output layer, which contains seven neurons corresponding to the seven different emotions in the dataset. All the layers except the output layer use the Exponential Linear Unit (ELU) activation function. The SoftMax activation function is used in the output layer to find the probability score of each emotion and to predict the emotions of the speech signal. This architecture allows the model to learn complex patterns in the high-dimensional input data.

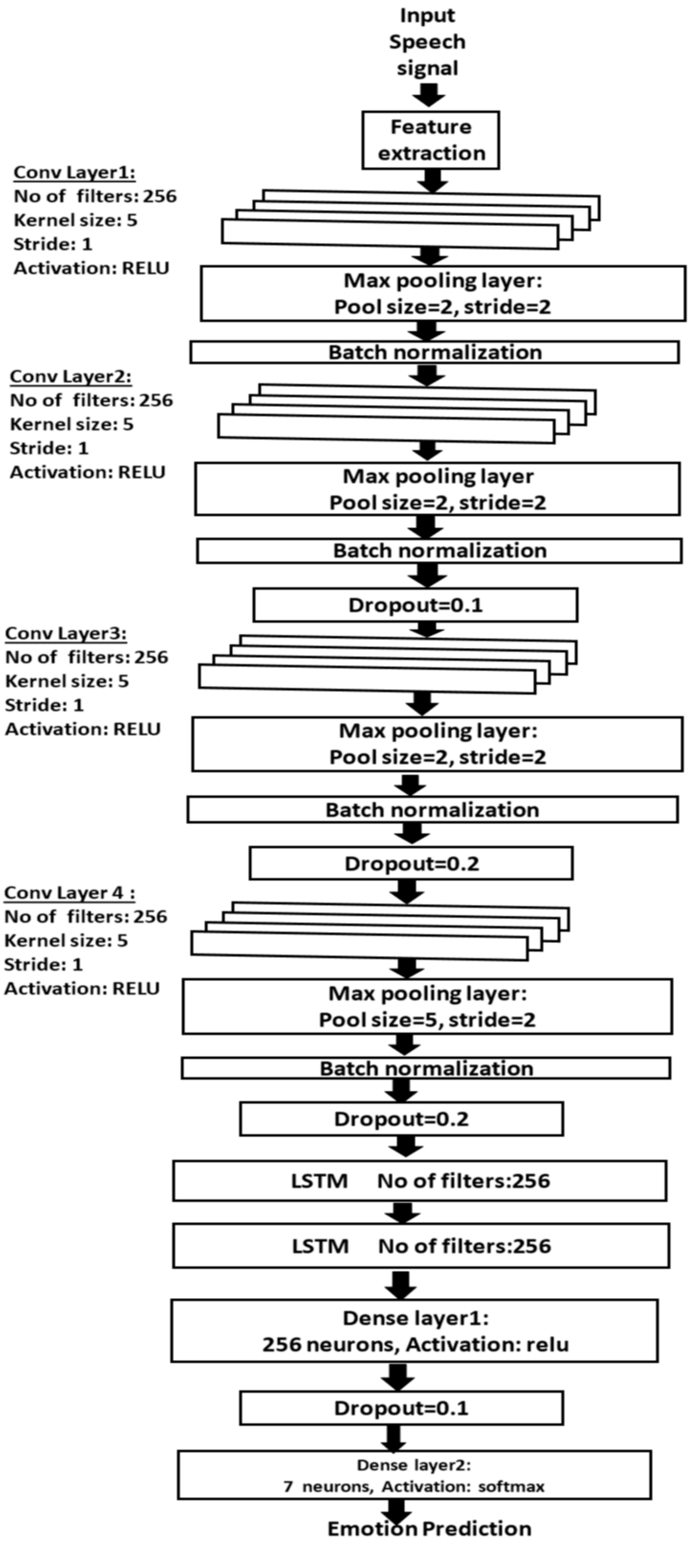


Figure 3.7: Proposed CNN+LSTM model for SER

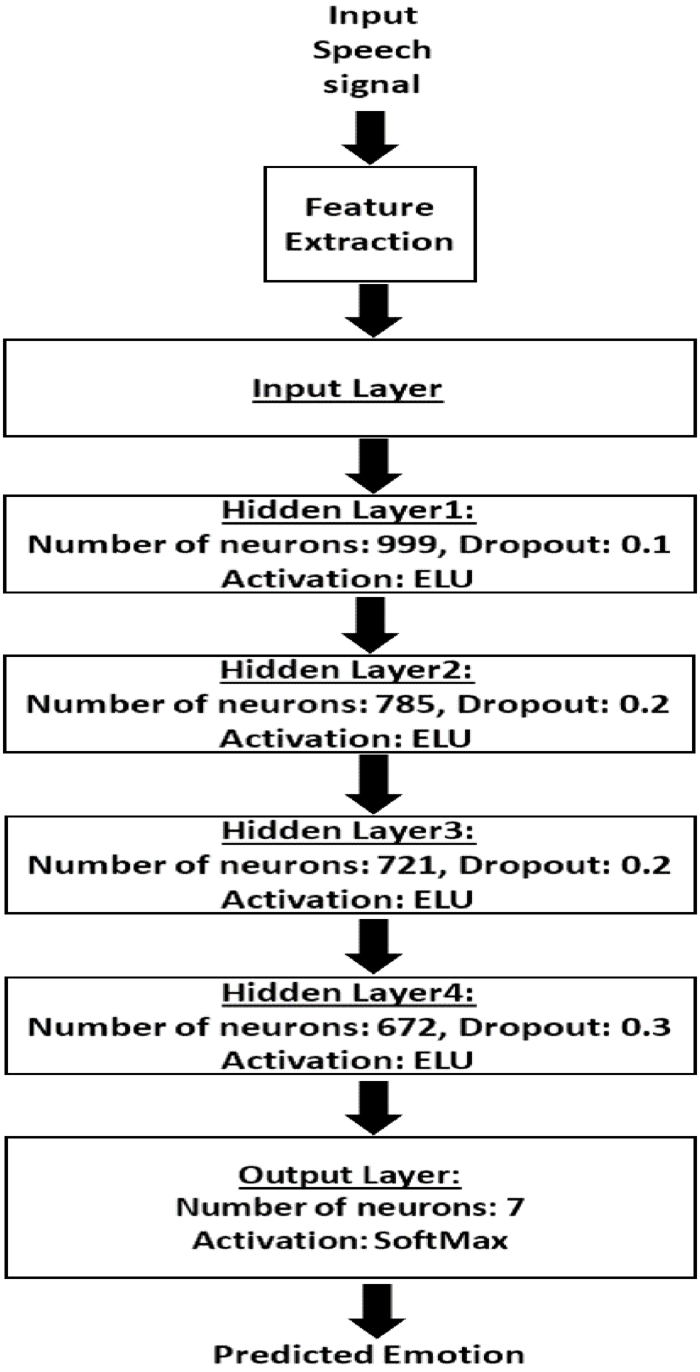


Figure 3.8: Proposed DNN model for SER

Chapter 4
Results And Analysis

In the current chapter, the results obtained after applying CNN+LSTM and DNN models for speech emotion recognition of the EmoDB dataset are discussed.

4.1 CNN+LSTM and DNN Using EmoDB

4.1.1 5-Fold Cross-Validation

1st iteration

In this case, Fold-1 is considered for test data, and the remaining folds are considered for training purposes. In this iteration, a validation accuracy of 76.63% was achieved using the CNN+LSTM model and a validation accuracy of 90.65% using the DNN model. The validation accuracy plots for Fold-1 using the CNN+LSTM and DNN models are displayed in Figures 4.1 and 4.2, respectively.

2nd iteration

In this case, Fold-2 is considered for test data, and the remaining folds are considered for training purposes. In this iteration, a validation accuracy of 75.70% was achieved using the CNN+LSTM model and a validation accuracy of 85.05% using the DNN model. The validation accuracy plots for Fold-1 using the CNN+LSTM and DNN models are displayed in Figures 4.3 and 4.4, respectively.

3rd iteration

In this case, Fold-3 is considered for test data, and the remaining folds are considered for training purposes. In this iteration, a validation accuracy of 73.83% was achieved using the CNN+LSTM model and a validation accuracy of 84.98% using the DNN model. The validation accuracy plots for Fold-1 using the CNN+LSTM and DNN models are displayed in Figures 4.5 and 4.6, respectively.

4th iteration

In this case, Fold-4 is considered for test data, and the remaining folds are considered for training purposes. In this iteration, a validation accuracy of 81.31% was achieved using the CNN+LSTM model and a validation accuracy of 85.98% using the DNN model.

The validation accuracy plots for Fold-1 using the CNN+LSTM and DNN models are displayed in Figures 4.7 and 4.8, respectively.

5th iteration

In this case, Fold-5 is considered for test data, and the remaining folds are considered for training purposes. In this iteration, a validation accuracy of 79.44% was achieved using the CNN+LSTM model and a validation accuracy of 88.79% using the DNN model. The validation accuracy plots for Fold-1 using the CNN+LSTM and DNN models are displayed in Figures 4.9 and 4.9, respectively.

The validation accuracy using a five-fold strategy is shown in Table 4.1. All five-fold accuracy plots are displayed above, revealing that the highest validation accuracy was 81% and the lowest testing accuracy was 73%. Using this 5-fold cross-validation technique, an average test accuracy of 77% was achieved with the CNN+LSTM model. The highest validation accuracy observed was 90%, and the lowest testing accuracy was 84%. Using the same 5-fold cross-validation approach, an average test accuracy of 87% was obtained with the DNN model.

Table 4.1: Accuracy of Five-Fold for EmoDB dataset

Fold	**Validation Accuracy in %**	
	CNN+LSTM	**DNN**
Fold-1	76.63	90.65
Fold-2	75.70	85.05
Fold-3	73.83	84.98
Fold-4	81.31	85.98
Fold-5	79.44	88.79
avergae of 5 folds	77.38	87.09

4.1.2 Using train-test-split

In this approach, train-test-split with a test size of 0.2 was employed to divide the dataset for training and testing. A validation accuracy of 77.57% was achieved using the CNN+LSTM model, and a validation accuracy of 85.98% was obtained using the DNN model.

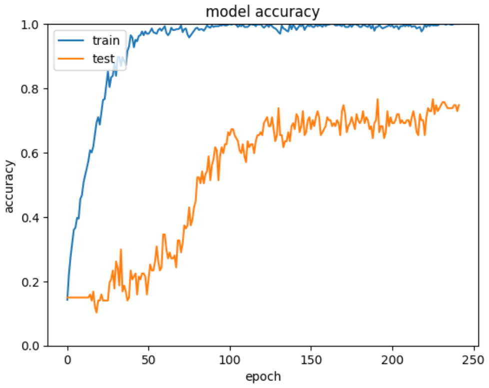


Figure 4.1: validation accuracy plot of Fold-1 using CNN+LSTM model

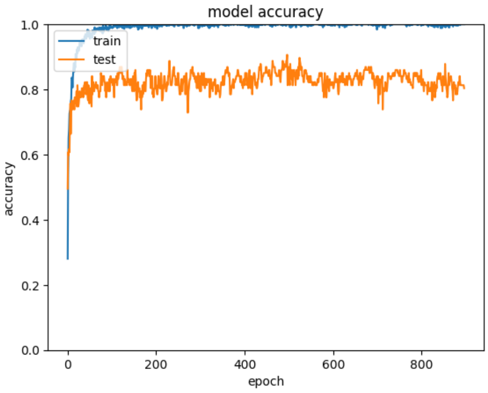


Figure 4.2: validation accuracy plot of Fold-1 using DNN model

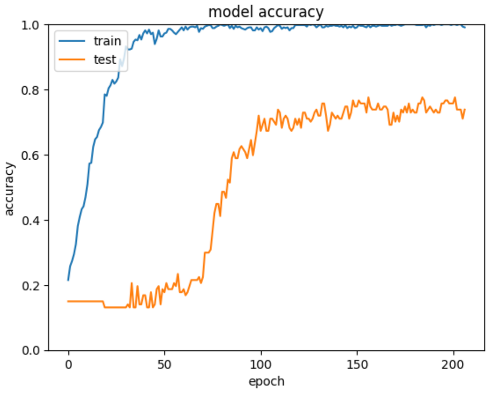


Figure 4.3: validation accuracy plot of Fold-2 using CNN+LSTM model

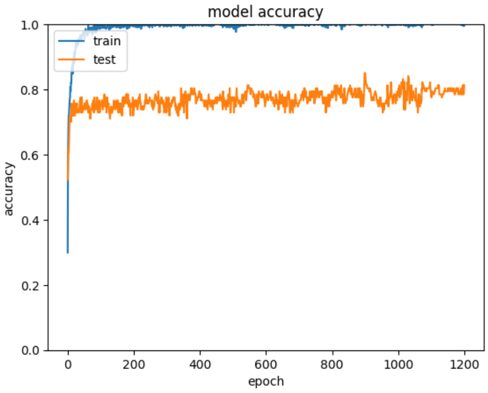


Figure 4.4: validation accuracy plot of Fold-2 using DNN model

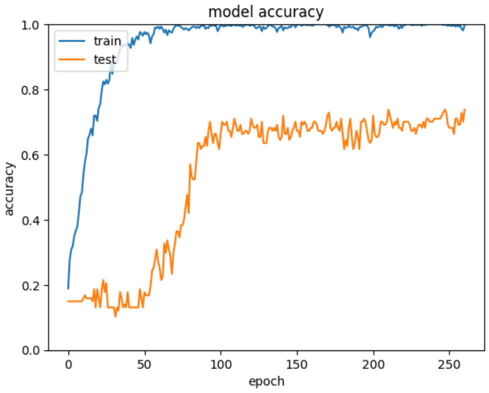


Figure 4.5: validation accuracy plot of Fold-3 using CNN+LSTM model

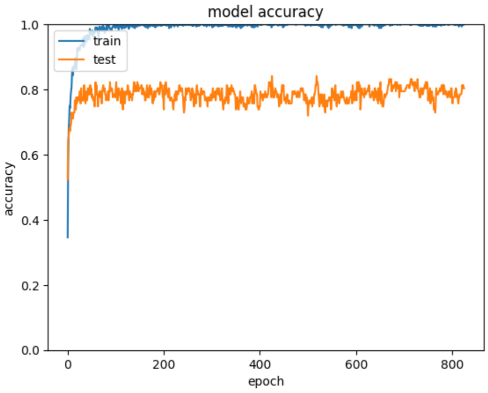


Figure 4.6: validation accuracy plot of Fold-3 using DNN model

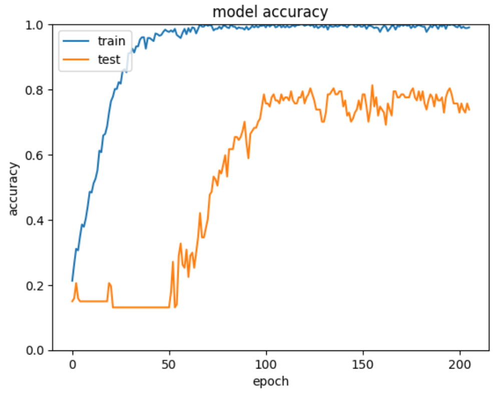


Figure 4.7: validation accuracy plot of Fold-4 using CNN+LSTM model

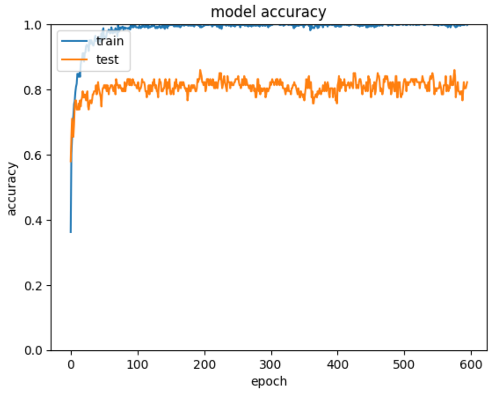


Figure 4.8: validation accuracy plot of Fold-4 using DNN model

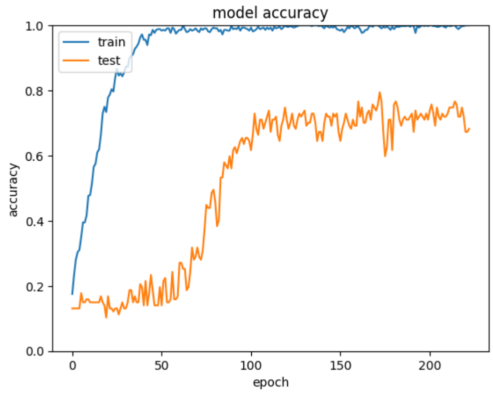


Figure 4.9: validation accuracy plot of Fold-5 using CNN+LSTM model

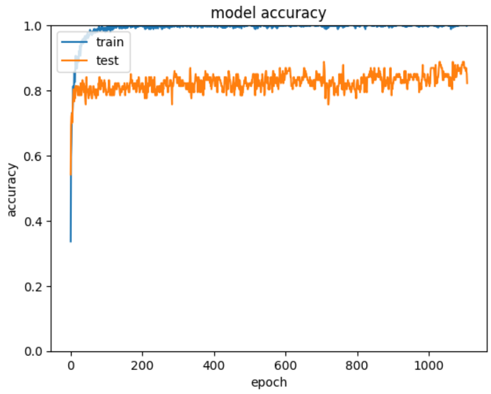


Figure 4.10: validation accuracy plot of Fold-5 using DNN model

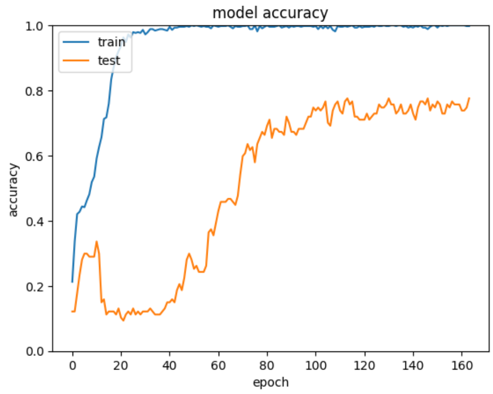


Figure 4.11: validation accuracy plot of train-test-split using CNN+LSTM model

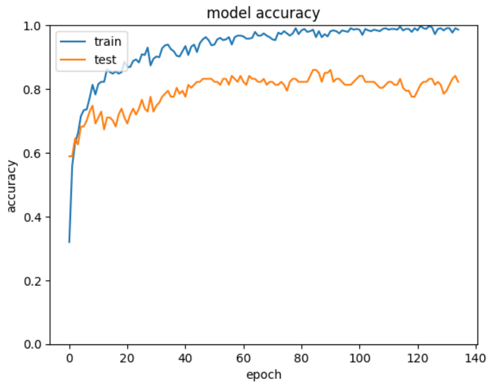


Figure 4.12: validation accuracy plot of train-test-split using DNN model

4.1.3 Classification report for train test split

A classification report is a valuable tool used to measure the quality of predictions from a classification algorithm. It provides a detailed breakdown of how well a model performs for each class in a classification problem. The Classification report for the EmoDB dataset using the proposed features and two models is shown in Tables 4.2 and 4.3 below. The classification report observed that emotions like $anger$, $disgust$, and $sadness$ are recognised with the highest recognition rate of 80%, 80%, and 92%, respectively. Emotions like $angry$, $boredom$, and $disgust$ were recognised with the highest recognition rate of 94%, 93%, and 91%, respectively.

Table 4.2: Classification report for CNN+LSTM model using EmoDB dataset

	precision	**recall**	**f1-score**	**support**
anger	0.80	0.84	0.82	19
boredom	0.71	0.83	0.77	18
disgust	0.80	0.73	0.76	11
fear	0.80	0.86	0.83	14
happiness	0.71	0.62	0.67	16
neutral	0.73	0.69	0.71	16
sadness	0.92	0.85	0.89	13
accuracy			0.78	107
macro avg	0.78	0.77	0.78	107
weighted avg	0.78	0.78	0.77	107
Average classification accuracy=77.57%				

Table 4.3: Classification report for DNN model using EmoDB dataset

	precision	**recall**	**f1-score**	**support**
anger	0.94	0.79	0.86	19
boredom	0.93	0.78	0.85	18
disgust	0.91	0.91	0.91	11
fear	0.82	1.00	0.90	14
happiness	0.80	0.75	0.77	16
neutral	0.79	0.94	0.86	16
sadness	0.86	0.92	0.89	13
accuracy			0.86	107
macro avg	0.86	0.87	0.86	107
weighted avg	0.87	0.86	0.86	107
Average classification accuracy=86%				

4.1.4 Comparison with the state-of-the-art models

In this section, compared proposed model results with results from state-of-the-art methods. Table 4.4 compares my performance with the state-of-the-art methods for the EmoDB dataset.

Table 4.4: PERFORMANCE COMPARISON OF SER WITH STATE-OF THE-ART METHODS ON THE EmoDB DATASET

Author	**Feature**	**Model**	**Accuracy**
Lukose et al. [35]	MFCC	GMM	76.31%
Liu et al. [36]	Formants	SVM+RBF	78.66%
Ancilin et al. [37]	MFMC	SVM	81.5%
Ozseven et al. [38]	MFCC, pitch, Formant, band width	SVM	82.8%
Proposed	**MFCC, Mel spectrogram, spectrogram, delta of MFCCs, chroma, tonnetz**	**CNN+LSTM**	**77.57%**
Proposed	**MFCC, Mel spectrogram, spectrogram, delta of MFCCs, chroma, tonnetz**	**DNN**	**86%**

Chapter 5
Conclusion & Future Scope

In this concluding chapter, the obtained results from the application of CNN+LSTM and DNN models for speech emotion recognition using the EmoDB dataset are summarized. Additionally, future research avenues are delineated to enhance the efficacy and applicability of speech emotion recognition systems.

5.1 Conclusion

The comparative study of the Deep Neural Network (DNN) model and the Convolutional Neural Network combined with the Long Short-Term Memory (CNN+LSTM) model, using a 5-fold cross-validation technique, revealed a significant difference in performance. The DNN model achieved an impressive average accuracy of 86% across the five folds, outperforming the CNN+LSTM model, which achieved an average accuracy of 76%. The feature extraction techniques employed, including Mel Spectrogram, MFCC, Delta MFCC, Tonnetz, and Chroma, were instrumental in achieving these results. The data for these models was sourced from a CSV file. This analysis underscores the effectiveness of the DNN model in this specific context, although the CNN+LSTM model also demonstrated a respectable performance. Future work could explore potential improvements to the CNN+LSTM model or investigate the impact of other feature extraction techniques.

5.2 Future Scope

Speech emotion recognition (SER) has gained significant traction in recent years, fueled by the transformative power of deep learning. This technology has revolutionised how we analyse and interpret human emotional cues conveyed through speech, enabling machines to comprehend and respond to the emotional nuances of human communication. With its ability to extract intricate patterns from complex data, deep learning has propelled SER to new heights of accuracy and robustness, paving the way for many promising applications in various domains.

Addressing Current Challenges

While SER technology has made remarkable strides, specific challenges must be addressed before widespread adoption can occur. One of the primary challenges lies in the sensitivity of SER systems to noise and variability in speech signals. To address this, researchers are exploring innovative techniques for noise reduction and developing signal-processing algorithms that can enhance the extraction of emotional information from speech.

Another challenge is the need for large-scale, high-quality datasets for SER. This limits the ability of deep learning models to effectively learn from diverse and representative data, potentially hindering their performance. To overcome this, researchers are actively working on creating and curating comprehensive SER datasets that capture a wide range of emotional expressions and contexts.

Emerging Applications

Numerous fascinating applications in various industries are emerging as SER technology matures. When it comes to customer service, SER systems can be used to automatically identify the feelings of customers, which allows companies to offer individualised and sympathetic assistance. This may improve customer satisfaction, retention, and general brand reputation. SER systems can track student participation in the classroom and spot possible emotional obstacles to learning. By using this information, educators can adjust their pedagogical strategies to meet the unique needs of each student better and create a more stimulating and productive learning environment. When diagnosing, planning treatments, and delivering patient-centred care, SER systems can be extremely helpful in gauging patients' emotional states. This may result in better patient outcomes, lower medical expenses, and higher patient satisfaction.

Beyond Human-Computer Interaction

Beyond human-computer interaction, SER has wider societal applications in its potential. SER systems, for example, can be used to evaluate social media sentiment and public opinion surveys, offering important insights into the emotional terrain of populations and communities. This data can help with crisis management plans, social interventions, and policy choices, resulting in better informed and efficient governance and societal well-being.

When it comes to mental health, SER systems can be used to identify early indicators of emotional distress and offer prompt support and intervention to those who are experiencing mental health issues. For those who are struggling with mental health issues, this may result in better mental health outcomes, lighter social loads, and an improved quality of life.

References

[1] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, “A review on speech emotion recognition using deep learning and attention mechanism,” *Electronics*, vol. 10, no. 10, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/10/1163

[2] S. Mishra, P. Warule, and S. Deb, “Speech emotion classification using feature-level and classifier-level fusion,” *Evolving Systems*, vol. 15, pp. 1–14, 11 2023.

[3] “DenseNet architecture, note=image, url=https://www.sharpsightlabs.com/blog/cross-validation-explained/,.”

[4] J. Baek and Y. Choi, “Deep neural network for predicting ore production by truck-haulage systems in open-pit mines,” *Applied Sciences*, vol. 10, p. 1657, 03 2020.

[5] R. Parameshwara, S. Narayana, P. Murugappan, R. Subramanian, I. Radwan, and R. Goecke, “Automated parkinson’s disease detection and affective analysis from emotional eeg signals,” 02 2022.

[6] C. Jiang, Y. Chen, S. Chen, Y. Bo, W. Li, T. Wenxin, and J. Guo, “A mixed deep recurrent neural network for mems gyroscope noise suppressing,” *Electronics*, vol. 8, p. 181, 02 2019.

[7] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639311000185

[8] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, “Sparse autoencoder-based feature transfer learning for speech emotion recognition,” 09 2013, pp. 511–516.

[9] F. Eyben, F. Weninger, F. Gross, and B. Schuller, “Recent developments in opensmile, the munich open-source multimedia feature extractor,” 10 2013, pp. 835–838.

[10] K. Han, D. Yu, and I. Tashev, “Speech emotion recognition using deep neural network and extreme learning machine,” 09 2014.

[11] S. Kim, S. J. Guy, K. Hillesland, B. Zafar, A. A.-A. Gutub, and D. Manocha, “Velocity-based modeling of physical interactions in dense crowds,” *The Visual Computer*, vol. 31, pp. 541–555, 2015.

[12] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE transactions on multimedia*, vol. 16, no. 8, pp. 2203–2213, 2014.

[13] D. Bertero, F. Siddique, C.-S. Wu, Y. Wan, R. Chan, and P. Fung, "Real-time speech emotion and sentiment recognition for interactive dialogue systems," 01 2016, pp. 1042–1047.

[14] A. Yadav and D. K. Vishwakarma, "A multilingual framework of cnn and bi-lstm for emotion classification," in *2020 11th international conference on computing, communication and networking technologies (ICCCNT)*. IEEE, 2020, pp. 1–6.

[15] G. Wen, H. Li, J. Huang, D. Li, E. Xun *et al.*, "Random deep belief networks for recognizing emotions from speech signals," *Computational intelligence and neuroscience*, vol. 2017, 2017.

[16] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69–78, 2019.

[17] M. Xu, F. Zhang, X. Cui, and W. Zhang, "Speech emotion recognition with multiscale area attention and data augmentation," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6319–6323.

[18] A. Mujaddidurrahman, F. Ernawan, A. Wibowo, E. A. Sarwoko, A. Sugiharto, and M. D. R. Wahyudi, "Speech emotion recognition using 2d-cnn with data augmentation," in *2021 International Conference on Software Engineering Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*, 2021, pp. 685–689.

[19] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, pp. 137–146, 2011.

[20] K. Aida-zade, A. Xocayev, and S. Rustamov, "Speech recognition using support vector machines," in *2016 IEEE 10th international conference on application of information and communication technologies (AICT)*. IEEE, 2016, pp. 1–4.

[21] D. Anguita, A. Ghio, S. Ridella, and D. Sterpi, "K-fold cross validation for error rate estimate in support vector machines." in *DMIN*, 2009, pp. 291–297.

[22] A. B. Ingale and D. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 1, pp. 235–238, 2012.

[23] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, “Speech emotion recognition using deep learning techniques: A review,” *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019.

[24] “Language-independent hyperparameter optimization based speech emotion,” 2022. [Online]. Available: https://link.springer.com/article/10.1007/s41870-022-00996-9

[25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, “Iemocap: Interactive emotional dyadic motion capture database,” *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[26] B. Schuller, S. Steidl, A. Batliner, P. Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E. Messner, K. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, “The interspeech 2018 computational paralinguistics challenge: Atypical and self-assessed affect, crying and heart beats,” 09 2018, pp. 122–126.

[27] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “Deap: A database for emotion analysis using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3, pp. 18–31, 12 2011.

[28] D. H. Rudd *et al.*, “Emodb dataset (berlin database of emotional speech),” 2022. [Online]. Available: https://paperswithcode.com/dataset/emodb-dataset

[29] “Emodb dataset — kaggle,” 2024. [Online]. Available: https://www.kaggle.com/datasets/piyushagni5/berlin-database-of-emotional-speech-emodb

[30] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, “A tutorial on onset detection in music signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[31] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[32] C. Harte, “Towards automatic extraction of harmony information from music signals,” 01 2010.

[33] M. Neuwirth, D. Harasim, F. Moss, and M. Rohrmeier, “The annotated beethoven corpus (abc): A dataset of harmonic analyses of all beethoven string quartets,” *Frontiers in Digital Humanities*, vol. 5, 07 2018.

[34] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[35] S. Lukose and S. S. Upadhya, “Music player based on emotion recognition of voice signals,” in *2017 international conference on intelligent computing, instrumentation and control technologies (ICICICT)*. IEEE, 2017, pp. 1751–1754.

[36] Z.-T. Liu, A. Rehman, M. Wu, W.-H. Cao, and M. Hao, “Speech emotion recognition based on formant characteristics feature extraction and phoneme type convergence,” *Information Sciences*, vol. 563, pp. 309–325, 2021.

[37] J. Ancilin and A. Milton, “Improved speech emotion recognition with mel frequency magnitude coefficient,” *Applied Acoustics*, vol. 179, p. 108046, 2021.

[38] T. Özseven, “Investigation of the effect of spectrogram images and different texture analysis methods on speech emotion recognition,” *Applied Acoustics*, vol. 142, pp. 70–77, 2018.