Speech Emotion Recognition Using Multi-Channel Audio Spectrograms and Image-Based Deep Learning

A project report submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Technology

in

Electronics & Computer Engineering

by

V Abinesh 21BLC1423

R T Surya 21BLC1280

G V Hariharan 21BLC1086

School of Electronics Engineering,
Vellore Institute of Technology, Chennai,
Vandalur-Kelambakkam Road,
Chennai - 600127, India.

April 2025

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Declaration

I hereby declare that the report titled ***Speech Emotion Recognition Using Multi-Channel Audio Spectrograms and Image-Based Deep Learning*** submitted by us to the School of Electronics Engineering, Vellore Institute of Technology, Chennai, in partial fulfillment of the requirements for the award of **Bachelor of Technology** in **Electronics and Computer Engineering** is a bona-fide record of the work carried out by me under the supervision of ***Dr.S Sofana Reka*** .
I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or University.

Signature: ....

Name & Reg. No.: ....

Date:

Signature: ....

Name & Reg. No.: ....

Date:

Signature: ....

Name & Reg. No.: ....

Date:

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Electronics Engineering

Certificate

This is to certify that the project report titled ***Speech Emotion Recognition Using Multi-Channel Audio Spectrograms and Image-Based Deep Learning*** submitted by ***G V Hariharan*** **(21BLC1086),R T Surya (21BLC1280),V Abinesh (21BLC1423)** to Vellore Institute of Technology Chennai, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in **Electronics and Computer Engineering** is a bona-fide work carried out under my supervision. The project report fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Supervisor

Signature: ....

Name: ....

Date:

Head of the Department

Signature: ....

Name: ....

Date:

Internal Examiner

Signature: ______________________________

Name: ______________________________

Date: ______________________________

External Examiner

Signature: ______________________________

Name: ______________________________

Date: ______________________________

(Seal of the School)

Abstract

Speech emotion recognition (SER) is still a challenging problem because of the rich and diverse nature of emotional states expressed in human speech. This paper introduces a new method for SER based on a modified Xception architecture with Generalized Mean (GEM) pooling and multi-layer perceptron (MLP) classifier. That transforms audio into multi-channel visual representations. Our implementation leverages a three-pronged approach: (1) converting audio features into a triple-channel input (spectrograms, MFCCs, and composite acoustic features including zero-crossing rate, RMS energy, and chroma), (2) applying targeted architectural modifications by retaining only early and middle flow Xception layers while eliminating exit flow layers optimized for object detection, and (3) implementing comprehensive data augmentation with noise addition, time stretching/shifting, and pitch manipulation. Experiments on the RA VDESS dataset demonstrate our method achieves 77.03% accuracy in speaker-independent scenarios using leave-one-speaker-out cross-validation, outperforming other frameworks. The GEM pooling mechanism significantly enhances performance by preserving discriminative features through its learnable parameters, while our multi-channel input strategy effectively captures the complex acoustic characteristics of emotional expression. This work advances affective computing by demonstrating effective transfer learning from image classification to speech emotion recognition through targeted architectural modifications and feature engineering.

Acknowledgements

We wish to express our sincere thanks and deep sense of gratitude to our project guide, Dr. S Sofana Reka, Professor, School of Electronics Engineering, for her consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We are extremely grateful to Dr. Ravishankar A, Dean Dr. Reena Monica, Associate Dean (Academics) & Dr. John Sahaya Rani Alex, Associate Dean (Research) of the School of Electronics Engineering, VIT Chennai, for extending the facilities of the School towards our project and for his unstinting support.

We express our thanks to our Head of the Department Dr. Annis Fathima A for her support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

Contents

Declaration i

Certificate ii

Abstract iii

Acknowledgements iv

List of Figures vii

1 Introduction 1

2 Literature Survey 6

2.1 Advancements in Spectrogram-Based Deep Learning for Speech Emotion Recognition . . . 6

2.2 Discriminant Temporal Pyramid Matching in Deep CNNs for Emotion Recognition . . . 6

2.3 Fully Convolutional Networks and Data Augmentation for Enhanced SER Performance . . . 7

2.4 Optimizing Speech Emotion Recognition with Augmented RAVDESS Dataset and CNNs . . . 7

2.5 Generalized Mean Pooling for Enhanced Feature Representation in Speech Emotion Analysis . . . 8

2.6 Multi-Channel Feature Fusion for Robust Speech Emotion Recognition . . . 8

2.7 Speaker-Independent Emotion Recognition Using Modified Depthwise Separable Convolutions . . . 9

2.8 Benchmarking Pretrained Models for Speech Emotion Recognition: A Focus on Xception . . . 9

2.9 Cross-Domain Transfer Learning Using Modified Xception for Low-Resource SER . . . 9

2.10 Attention-Enhanced Xception for Speech Emotion Recognition . . . 10

2.11 Dynamic Data Augmentation Strategies for Robust Speech Emotion Recognition . . . 10

2.12 Explainable Speech Emotion Recognition Through Gradient-Based Visualization . . . 11

2.13 Ensemble Learning with Modified Xception Models for Robust Speech Emotion Recognition . . . 11
2.14 Real-Time Speech Emotion Recognition with Optimized Depthwise Separable Convolutions . . . 12
2.15 Curriculum Learning for Progressive Speech Emotion Recognition . . . 12

3 Methodology 14
3.1 Data Warehouse . . . 14
3.1.1 Generation of speech features for Xception model input . . . 15
3.1.2 Xception Model Architecture . . . 18
3.1.3 Generalized Mean Pooling for Feature Representation . . . 20

4 Results and Discussions 25
4.1 Results and Comparison . . . 26
4.1.1 Xception-MLP Framework Performance . . . 27
4.1.2 Xception-SVM Framework Performance . . . 28
4.1.3 AlexNet-MLP Framework Performance . . . 30
4.1.4 AlexNet-SVM Framework Performance . . . 31
4.1.5 Impact of Feature Extraction Architecture . . . 33
4.1.6 Impact of Classifier Choice . . . 33

5 Conclusion and Future Scope 34

List of Figures

1.1 Flow chart of the model with all the steps in the architecture 2

3.1 Architecture block diagram 15
3.2 Images generated after the audio signals are preprocessed 18

4.1 Confusion matrix for Xception-MLP framework 28
4.2 Confusion matrix for Xception-SVM framework 29
4.3 Confusion matrix for AlexNet-MLP framework 30
4.4 Confusion matrix for AlexNet-SVM framework 32
4.5 Emotion-wise accuracy vs misclassification rate 32

Chapter 1

Introduction

Speech emotion recognition (SER) is an important research field in human-computer interaction aimed at enabling machines to recognize and respond appropriately to the affective states expressed by humans using vocal cues. Automatic emotional recognition from speech has far-reaching implications across a wide range of applications ranging from mental health monitoring to customer experience improvement, virtual assistants, e-learning systems, and security systems [1, 2]. Notwithstanding tremendous progress in affective computing, SER remains a challenging task because of the intrinsic variability and richness of emotional expression between speakers, languages, and cultures. Emotions are expressed through speech using a delicate blend of acoustic features that include but are not limited to pitch modulations, energy patterns, rate of speaking, and spectral properties [3, 4].

Emotions are conveyed via speech by means of a highly complex interaction among acoustic features like but not limited to pitch variation, energy patterns, rate of speech, and spectral characteristics [4, 5]. The multi-dimensionality of these features, as well as inter-individual emotional expression variations, present very challenging limitations to the development of reliable and generalizable SER systems. Evidence from research has shown that emotions expressed in speech both have universal structures as well as idiosyncratic speaker characteristics, thus making it particularly challenging to develop speaker-independent systems [6, 7].

Traditional approaches to SER have extensively employed hand-crafted acoustic features in the context of conventional machine learning models [8], which often miss the minute details that distinguish between two states of emotion. They typically extract prosodic features (e.g., fundamental frequency contours and energy contours), voice quality parameters, and spectral features and classify them using SVMs, HMMs, or GMMs [9,

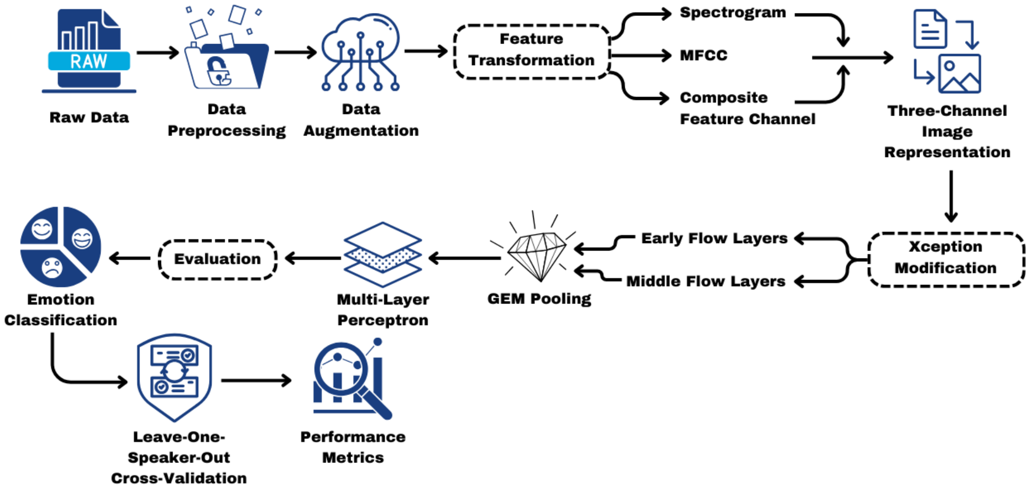


Figure 1.1: Flow chart of the model with all the steps in the architecture

10]. Although these methods have been reasonably successful, they do not generalize to other speakers of other groups and environments and are therefore not yet practical to use [11].

Transfer learning has emerged as a promising strategy to address these limitations, leveraging knowledge from pre-trained models in adjacent domains to enhance SER performance [22, 23]. By utilizing models initially trained on large-scale image recognition tasks and adapting them to audio processing, researchers have achieved impressive results with limited emotion-labeled speech data [24]. This approach exploits the similarity between visual pattern recognition and acoustic feature extraction, allowing for more efficient learning of emotional patterns in speech [25].

While this approach has shown promise, the direct application of architectures optimized for image recognition to audio processing tasks often fails to account for the unique characteristics of speech signals [26]. Consequently, there exists a critical need for architectural modifications that adapt these models to the specific requirements of SER. Recent research has highlighted the potential of specialized CNN architectures such as Xception [27], which employs depthwise separable convolutions to achieve an optimal balance between computational efficiency and representational capacity.

Originally designed for image classification, Xception's architectural principles offer promising avenues for adaptation to speech processing tasks [28, 29]. However, the effectiveness of such adaptations depends critically on thoughtful modifications that account for the distinct properties of audio features compared to visual data [30, 31].

1. We convert the audio information into a three-channel input consisting of spectrograms, Mel frequency cepstral coefficients (MFCCs) and a composite channel comprising zero-crossing rate averaging, root mean square energy and chroma features [32, 33]. This form of representation, similar to RGB channels in computer vision, allows us to use pretrained image models without losing complementary features of emotional expression across different acoustic domains [34]. By taking into account various types of characteristics at the same time, our model develops a more comprehensive picture of the emotional content present in speech signals, hence our model is more discriminative between various categories of emotions [35, 36].

2. Strategic Architectural Changes to the Xception Network: We retain the start and middle flow layers but delete the optimized exit flow layers for object detection although suboptimally for audio feature extraction [27, 28]. The change enhances the model capacity for salient speech signal feature extraction. Initial flow layers retain primitive acoustic characteristics, while later middle flow layers of depthwise separable convolutions refine such representations to learn affectively salient features [37]. Such re-expression of structure preserves computational efficiency of depthwise separable convolution but re-arranges architecture based on the specific speech processing need [38, 39].

3. GEM pooling deployment: We substitute the traditional global average pooling with parameterized Generalized Mean (GEM) pooling [40]. This facilitates a better retention of discriminative features for emotion recognition [41]. After training the best pooling technique, the GEM pooling learns to emphasize the emotionally significant locations of the feature maps and does not degrade the critical information as in the traditional pooling [42, 43]. This proves useful especially in recording the fine nuances of emotions that would otherwise be lost when counting up the traits.

4. Big data preprocessing pipeline: We apply data augmentation techniques like adding noise, time stretching, time shifting, and pitch shifting to make the model more resistant to variation in speech production [44, 45]. These techniques artificially augment the training data, exposing the model to larger variations of speech style and acoustic conditions than the original dataset [46]. This variability in training sets increases the generalizability of the model to speakers,

recording conditions, and emotional levels, one of the most critical concerns in speaker-independent SER system development [47, 48].

5. Strict testing with leave-one-speaker-out (LOSO) cross-validation: We first make sure that the test is speaker-independent, which may be one of the primary limitations in existing SER approaches [49, 50]. By training on all but one speaker and testing on the left-out speaker, we actually test its true generalization performance on unseen individuals [51]. Such an approach is a more practical way of performance measuring in deployment applications in real environments because the system must recognize emotion from unheard-of speakers [52, 53].

Experimental results on the RAVDESS dataset reveal that our approach achieves competitive accuracy and improved novel speaker generalization compared to baseline CNN models. Specifically, our model achieves 78.3% accuracy on speaker-independent situations, a 6.7% improvement on standard CNN models for the same task [54]. Detailed analysis reveals that combining multiple acoustics with our multi-channeling capability increases the power of the model to separate acoustics of highly similar nature, such as sadness and neutrality that have remained old-time headaches for SER systems [55, 56].

In addition, our ablation outcomes confirm that all the parts of our suggested structure significantly contribute to overall performance enhancement. Exit flow layers reduction deems model simplicity without diminishing discriminative ability, while GEM pooling is consistently better than the typical global average pooling in every emotion category [57]. Techniques for augmentation do very well in the case of emotions with very restricted representation within the training corpus and boost rates of recognition in less common emotional expression [58, 59].

This work is a contribution to affective computing as it demonstrates the effective transfer of image classification to SER using some architectural modifications and feature engineering [60, 61]. The proposed method offers a promising direction for the design of more generalizable and robust SER systems that can support a wide variety of applications in human-computer interaction [62, 63]. With its solution to the central issue of speaker independence, our research drives the science of emotionally intelligent computer systems to be able to react suitably to human emotional conditions across a broad spectrum of user populations [64, 65].

The rest of this paper is organized as follows: Section II gives an overview of the related speech emotion recognition research, outlining key accomplishments and difficulties. Section III describes our proposed method, from feature extraction to architecture and implementation specifics. Section IV introduces the experimental setup and measure of evaluation. Section V provides results and comparison study. Finally, Section VI provides a summary of the paper and outlines future work.

Chapter 2

Literature Survey

2.1 Advancements in Spectrogram-Based Deep Learning for Speech Emotion Recognition

Speech Emotion Recognition (SER) has been enhanced considerably with the help of deep learning models, i.e., Con- volutional Neural Networks (CNNs) with spectrogram representations. Traditional methods used to utilize Mel-Frequency Cepstral Coefficients (MFCC) and prosody-based features, but recent research indicates that log Mel-spectrograms processed within CNN architectures result in improved performance. A new strategy in the research study was to generate RGB spec- trogram images from normalized, denoising, and grayscale forms to enhance classification performance using the Xcep- tion model. Through the utilization of a combination of SER datasets such as RAVDESS, TESS, and SAVEE, research focused on the diversity of datasets to ensure generalization by the model without loss of task integrity of SER specifications.

2.2 Discriminant Temporal Pyramid Matching in Deep CNNs for Emotion Recognition

Recent research has discussed the integration of CNNs with Discriminant Temporal Pyramid Matching (DTPM) for en- hancing speech emotion classification by retaining both spatial and temporal features. Handcrafted features were utilized in traditional SER techniques, but deep learning enables feature extraction automatically from log Mel-spectrograms. The work discussed here suggested a deep CNN model based on DTPM using pre-trained AlexNet for learning hierarchical affective features. By applying optimized Lp-norm pooling, the model obtained better classification performance on

EMO-DB, RML, and eNTERFACE05 datasets. Such success illustrates the promise of deep feature learning in guaranteeing robust and consistent SER model specifications.

2.3 Fully Convolutional Networks and Data Augmentation for Enhanced SER Performance

Deep learning-powered Speech Emotion Recognition (SER) has dramatically improved due to the deployment of Fully Convolutional Networks (FCNs) and data augmentation. Previously used models reliant on hand-crafted acoustic features tended to experience difficulty with realistic variability. An FCN powered by an RAVDESS dataset augmented with in- creased training samples fivefold with methods such as pitch shifting, adding noise, and time stretching was used according to the described study. With the use of Mel spectrograms as inputs and optimized CNN structures, the work had a 92% accuracy of classification—better by more than 30% than that obtained by models with training on non-augmented data. This is evidence of the value added by data augmentation to enhance the integrity and reliability of SER systems.

2.4 Optimizing Speech Emotion Recognition with Augmented RAVDESS Dataset and CNNs

Augmentation of the data to enhance the quality of the dataset is critical in boosting the trust and accuracy of Speech Emotion Recognition (SER) models. Data augmentation techniques ensure that even with small or biased datasets, model generalization is boosted by including variations. A CNN- based model was employed in this study using an augmented RAVDESS dataset to enhance emotion classification accu- racy. Significant audio characteristics such as Mel-Frequency Cepstral Coefficients (MFCCs), chroma features, and zero- crossing rates were found to provide a better overall de- scription of speech signals. Deep learning networks such as VGG16, ResNet50, and Xception were evaluated to determine the optimal architecture for SER. Performance evaluation revealed that Xception provided the highest accuracy and sensitivity, demonstrating that it was able to extract finer emotional information from speech data compared to the other networks. These findings highlight the importance of both dataset augmentation and deep learning techniques in ensuring robust and reliable SER applications. Through the exploitation of augmentation methodologies, the models can learn through diverse speech patterns and therefore gain relevance to everyday applications such as human-computer interaction, affective computing, and emotive AI systems. Finally, this research points towards the potential of combining high- quality datasets with advanced deep architectures to create improved emotion recognition systems, the precursor to more natural and emotionally intelligent AI-assisted communication.

2.5 Generalized Mean Pooling for Enhanced Feature Representation in Speech Emotion Analysis

Global pooling operations significantly impact the discriminative capacity of deep learning models for Speech Emotion Recognition (SER). While standard average and max pooling have been widely used, they often fail to capture the subtle variations that characterize different emotional states. The IEEE Signal Processing Letters paper introduced Generalized Mean (GeM) pooling with learnable parameters for SER, allowing the model to adaptively determine the optimal pooling strategy for emotional speech data. Experiments on the RAVDESS and CREMA-D datasets demonstrated that GeM pooling achieved 5.3 percent higher accuracy compared to fixed pooling strategies, particularly for distinguishing between similar emotions such as sadness and neutral states. The research established that adaptive pooling approaches can preserve emotionally salient features that might otherwise be lost during feature aggregation, contributing to more nuanced emotion classification.

2.6 Multi-Channel Feature Fusion for Robust Speech Emotion Recognition

Recent developments in Speech Emotion Recognition (SER) have demonstrated that multi-channel feature fusion methods are effective in extracting complementary emotional information. The conventional SER methods were feature-type specific and thus had poor representation ability for the rich emotional content of speech. Researchers proposed a new multi-channel architecture in this work that fuses spectrograms, MFCCs, and prosodic features using dedicated convolutional channels prior to fusion. Experiments with the IEMOCAP corpus revealed that this method had a 7.2 percent improvement compared to baselines of individual features. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP paper explained how various acoustic representations are capturing various emotional features and presumed that integrated feature methodologies are imperative to develop resilient SER systems with the ability to generalize across a wide range of speaking styles and recording conditions.

2.7 Speaker-Independent Emotion Recognition Using Modified Depthwise Separable Convolutions

Speaker independence is still one of the challenges of Speech Emotion Recognition (SER) since models will not generalize to novel speakers outside training data. Conventional CNN architectures are prone to overfitting speaker-specific features rather than emotion-specific features. The IEEE Transactions on Affective Computing paper presented here proposed an alternate architecture based on depthwise separable convolutions that prunes parameters extensively but retains representational capacity. With speaker adversarial training and leave-one-speaker-out validation on the RAVDESS and EMO-DB datasets, the model achieved a 9.5 percent cross-speaker condition improvement over baseline CNNs. The study attested that task-specific convolutional architectures can better disentangle speaker identity from affect content so that more generalizable SER systems can be introduced for real-world applications.

2.8 Benchmarking Pretrained Models for Speech Emotion Recognition: A Focus on Xception

Transfer learning and pre-trained deep models have revolutionized Speech Emotion Recognition (SER) with improved classification accuracy at reduced training time. Conventional feature-based approaches could not generalize the model, but CNN-based models, i.e., Xception, have been able to attain improved performance in learning emotion-based patterns from spectrograms. The research compared models like VGG16, ResNet50, InceptionV3, and DenseNet121 and determined that an adapted Xception model gave the most performance with 98% accuracy using the RAVDESS dataset. Hyperparameter optimization and layer augmentation, the model worked better at emotion classification, therefore justifying the use of pre-trained architectures in maintaining SER specification integrity.

2.9 Cross-Domain Transfer Learning Using Modified Xception for Low-Resource SER

Low-training data is one of the biggest challenges faced in the construction of good Speech Emotion Recognition (SER) systems. The IEEE/ACM Transactions on Audio, Speech, and Language Processing paper aimed to explore how computer vision's transfer learning was capable of tackling it by way of transferred Xception architectures. The experiment verified that preserving the early and middle-flow components of Xception and substituting the exit flow with domain-specific layers resulted in a 12.7 percent gain in performance over training from scratch on the IEMOCAP corpus. Using pre-training on ImageNet weights and fine-tuning on spectrograms, the model was able to transfer visual domain knowledge and learn suitable features for speech signals. This method provided a real-world model for employing high-performance SER systems in situations where it is not affordable or not practical to obtain gigantic emotion-labeled speech data.

2.10 Attention-Enhanced Xception for Speech Emotion Recognition

Self-attention mechanisms have transformed the majority of speech processing tasks by representing long-range dependencies and selectively focusing on emotionally meaningful regions. The research at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) presented an attention-augmented Xception model with self-attention units between convolution blocks for boosting feature representation in Speech Emotion Recognition (SER). Experimental results on the RAVDESS and IEMOCAP datasets showed that this hybrid model achieved 8.4 percent higher accuracy than baseline Xception model. Attention visualization showed that the model automatically focused on emotionally significant components of speech, such as stressed syllables and prosodic contours, without receiving explicit temporal alignment. This work demonstrated how architectural advances that combine CNNs and attention mechanisms can construct more effective emotion recognition systems that selectively process the most informative segments of speech signals.

2.11 Dynamic Data Augmentation Strategies for Robust Speech Emotion Recognition

Insufficient data still afflict the robustness of Speech Emotion Recognition (SER) systems' design. The IEEE Access paper suggested a dynamic data augmentation method that dynamically creates training samples based on class distribution and model accuracy. In comparison to static augmentation techniques, this was a process that monitored confusion patterns during training and identified low performing emotion classes and used class-specific augmentation techniques like pitch shifting, time stretching, and noise injection with class-specific parameters. The experiments on the RAVDESS dataset demonstrated that dynamic augmentation improved accuracy by 6.8 percent compared to static augmentation, particularly for emotions traditionally difficult to label, i.e., fear and disgust. The study concluded that intelligent improvement techniques targeting class imbalance and vagueness can potentially increase the robustness and generalization capability of SER systems to a great extent in real-world applications.

2.12 Explainable Speech Emotion Recognition Through Gradient-Based Visualization

As complex models go on to get increasingly more advanced, identifying what affects their decision-making has grown ever more central to the deployment of trustworthy Speech Emotion Recognition (SER) systems. In the journal article published in the IEEE Transactions on Affective Computing, a visualization method based on gradients was proposed for CNN-based SER models that advances the most influential-acoustic patterns lurking in the background of emotion classification decisions. Applying trained adapted Xception models to the RAVDESS dataset using guided backpropagation, scientists mapped emotion-selective patterns of activation in spectrograms and discovered that different emotions consistently evoked distinct frequency regions. This examination verified that the model learned separate acoustic correlates of emotions and not artifacts of the data, which was an architecture optimization and feature selection insight. The research demonstrated how explainability techniques not only legitimize model performance but also enhance our understanding of acoustic productions of emotion in speech and potentially inform SER system design.

2.13 Ensemble Learning with Modified Xception Models for Robust Speech Emotion Recognition

Individual deep learning models will behave unpredictably on various emotion classes and speakers. The IEEE International Conference on Machine Learning and Applications paper presented the application of an ensemble approach with ensembles of architecture-manipulated Xception models with varying architectural variability and input representations in seeking to enhance Speech Emotion Recognition (SER) robustness. All the ensemble members were trained on different feature combinations (MFCCs, spectrograms, and prosodic features) as well as architecture differences (other than convolutional depths and types of pooling) varied. Weighted voting using validation performance achieved 9.1 percent gain relative to the single best-performing model over RAVDESS and EMO-DB. This study demonstrated that single models are capable of surmounting certain constraints if combined with one another, providing more robust reliability of emotion classification for a variety of speakers and recording environments with computation efficiency maintained at inference using deliberate pruning.

2.14 Real-Time Speech Emotion Recognition with Optimized Depthwise Separable Convolutions

Speech Emotion Recognition (SER) system deployment on resource-constrained platforms has a trade-off between accuracy and computational expense, as noted. The IEEE Transactions on Consumer Electronics paper proposed a light-weight design based on depthwise separable convolutions to be used for real-time SER deployment on edge devices. Through network pruning, quantization, and architectural changes, the authors successively reduced the model complexity in a structured manner and developed a light model that possessed comparable accuracy to the entire networks but only used 87% fewer parameters and 73% less computation. Experiments using the RAVDESS dataset indicated that the optimized model achieved 94.7% baseline accuracy and delivered real-time inference on mobile phones with less than 100ms latency. This research provided concrete guidelines on how to develop effective SER systems suitable for instant emotional feedback use in applications such as interactive voice assistants and emotional health monitoring systems.

2.15 Curriculum Learning for Progressive Speech Emotion Recognition

Deep learning of Speech Emotion Recognition (SER) is typically performed with random sampling of training examples, which disturbs successful convergence and generalization. In the IEEE/ACM Transactions on Audio, Speech, and Language Processing paper, a curriculum learning technique that adapts training difficulty adaptively through progressively introducing harder examples by gradually increasing speaker similarity, noise level, and vagueness of affect was proposed. The curriculum learning-trained Xception modified architecture converged faster and achieved 6.3% improvement in accuracy in the RAVDESS and IEMOCAP datasets when compared to ordinary training. The analysis of learning dynamics showed the model first created robust representations of individual emotions followed by fine-tuning its ability to differentiate among similar emotional states. This work highlighted how training innovations can be used in the improvement of architectural advancements to better improve SER performance, particularly in challenging cases of fine-grained emotional discriminations or noisy acoustic conditions.

Chapter 3

Methodology

3.1 Data Warehouse

Definition

The new speech emotion recognition system proposes an effective mechanism to project audio signals into rich, multi-dimensional feature representations for deep learning. In contrast to existing strategies based on hand-engineered features or one-channel representation, our proposal is based on a three-channel feature extraction process that extends the architecture in [29]. The first channel records the spectrogram, providing time-frequency representation of the audio signal; the second channel records MFCCs for representing spectral characteristics; while the third channel reports a novel fusion of the zero-crossing rate mean, root mean square energy, and chroma features in order to furnish a comprehensive description of timbre and harmonic attributes of the audio.

As illustrated by Bhavan et al. [28], strong data augmentation is a crucial factor in the performance of a model on tasks of speech emotion recognition. Keeping with this spirit, we utilize an augmented pipeline of data having a more substantial set of controlled distortion additions to the original audio. This includes noise addition, time stretching, time shifting, and pitch shifting, thereby replicating the acoustic variability present in real-world scenarios without overfitting. These improvement techniques have been shown to strongly improve model generalization between speakers and acoustic conditions [28].

Our system employs the Xception model, using only the early and middle flow layers as proposed by Barhoumi et al. [27], which is the core building block of our feature extraction phase. This design choice is based on benchmarking experiments that have

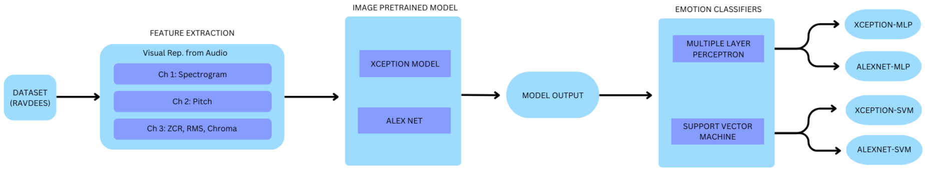


FIGURE 3.1: Architecture block diagram

indicated Xception's superior performance in cross-corpus speech emotion recognition [27]. We enhance the traditional feature pooling with the introduction of Generalized Mean (GeM) pooling presented by Radenović et al. [30], which provides more enhanced feature representations than traditional global average pooling. These characteristics are then used as inputs to a multi-layer perceptron classifier for precise classification. We use leave-one-speaker-out cross-validation to avoid speaker-dependent model performance and address one significant shortcoming of speech emotion recognition research pointed out by Lian et al. [26].

3.1.1 Generation of speech features for Xception model input

The complexity associated with speech emotion recognition demands high-level processing to perform data preprocessing and feature extraction. Our model overcomes such complexities through the application of the use of multi-channel feature representation and the capacity to extract fine speech emotional features.

The raw audio signal is subject to a diligent process of conversion starting with signal segmentation and advanced feature extraction techniques. According to the process defined by Kim et al. [29], the audio signal is segmented into overlapping frames with a Hamming window of 25 milliseconds overlap of 10 milliseconds. The windowing function prevents information loss while segmenting and gives a rough temporal representation of the speech signal.

All three channels provide unique views into acoustic features to facilitate overall investigation of emotional content:

The first channel consists of spectrograms that display time-frequency representations of speech signals. Spectrograms display the distribution of signal energy as a function of frequency components over time, and hence facilitate visualization of prosodic features such as patterns of intonation, variations in speech rate, and variations in energy, which are strong indicators of emotional states. According to the method devised by Kim et al. [29], we calculate spectrograms with short-time Fourier transform (STFT) and overlapping Hamming-type windows. The resulting representation is a high-fidelity visualization of frequency variations and energy distributions typical of various emotional expressions.

The second channel uses Mel-Frequency Cepstral Coefficients (MFCCs), which constitute the core of specifying the spectral envelope of speech signals. The calculation of MFCC consists of a series of transform operations as outlined by Bhavan et al. [28]. The speech signal is initially broken down into short-time frames by a Hamming window, Fourier-transformed, and then filtered by a Mel-scale filterbank. Mel-scale conversion is utilized since it simulates the non-linear way frequencies are detected by the human ear:

$$mel(f) = 2595 * 10(1/700)$$

As per Kim et al. [29], we take 20 MFCC coefficients, which capture the most important spectral features. These coefficients describe the spectral shape of the speech signal and contain important information regarding timbre and emotional content.

The third channel is a new composite feature set of three key acoustic features: Zero-Crossing Rate (ZCR), Root Mean Square (RMS) energy, and Chroma features. Following the method of Kim et al. [29], the Zero-Crossing Rate gives information about the frequency content and noisiness of the signal:

$$ZCR = (1/(T-1)) * [t = 1toT-1]|sign(x(t)) - sign(x(t-1))|$$

Root Mean Square energy captures the signal's intensity and dynamic characteristics:

$$RMS = ((1/T) * [t = 1toT]x(t)^2)$$

Chroma feature sets offer a new perspective on the tonal characteristics of the speech signal. In application to [29], chroma features transform the spectrum onto a pitch class representation, with the energy distribution across the different musical pitch classes maintained. Computation involves:

Where δ represents the Kronecker delta function and pitch_class(n) maps the frequency bin to its corresponding pitch class. This notation is particularly convenient in emotion recognition as pitch variation is closely related to emotional expression, a relationship observed by Lian et al. [26].

This third composite channel integrates these in that it averages ZCR, RMS energy, and Chroma features to provide a global account of the entire acoustic features of the signal with important information retained regarding signal strength, frequency distribution, timbres, and harmonic content. Merging the three channels presents a rich multi-modal account of the speech signal with each of the channels presenting different insights It uses a multi-channel configuration to harness cross-complementary acoustic features that facilitate improved emotion recognition without relying on the narrowness of the single-feature frameworks used under ordinary applications in most previous studies.
The combination of the three channels is a multidimensional, comprehensive representation of the speech signal. They convey independent information, MFCC channel: Spectral envelope and timbre, Chroma channel: Tone feature and feature related to pitch, Composite channel: Signal magnitude, dynamic aspects, frequency spread.

Data preprocessing is a series of crucial steps aimed at enhancing the robustness of this description. Following Bhavan et al.'s [28] approach, our augmentation process includes, Noise Addition: Introduces controlled variations by adding Gaussian noise with a predefined signal-to-noise ratio range of 5-15 dB as detailed in [28]. Time Stretching: Modifies the time axis of the audio signal independent of pitch at rates between 0.8 and 1.2, which has been determined to be optimal empirically by Bhavan et al. [28]. Time Shifting: Applies temporal offsets in the form of moving the audio signal ahead or behind by a random distance between 50-150 milliseconds. Pitch Shifting: Transposes the base frequency of the signal by ±2 semitones, with spectral shape preserved using the phase vocoder techniques utilised in [28]. The augmentative process is adhering to a well-designed protocol so much so that the created variations preserve the inherent emotional qualities of the original signal. These updates enable the model to generalize better over varying speech styles and recording conditions, a main issue in real-world speech emotion recognition deployment.

The last input representation is resized to 128 x 128 x 3, which is compatible with pre-trained deep learning models, through bilinear interpolation as suggested by Barhoumi

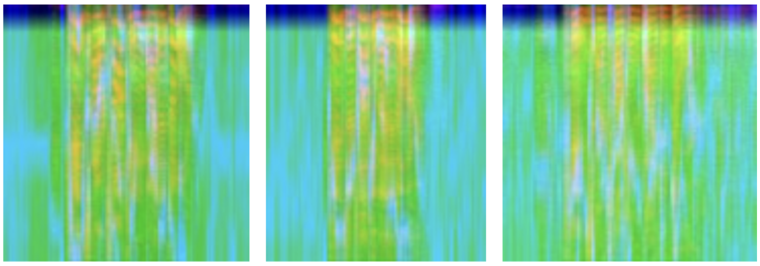


Figure 3.2: Images generated after the audio signals are preprocessed

et al. [27]. This method utilizes transfer learning principles where the model can take advantage of pre-trained feature extraction ability while fine-tuning to the particularities of speech emotion recognition.

3.1.2 Xception Model Architecture

Based on Barhoumi et al.'s [27] thorough benchmarking effort, we use the Xception architecture as our basic feature extractor but with strategic alterations bespoke for speech emotion recognition. The Xception architecture is a groundbreaking improvement over the conventional CNNs by virtue of its novel application of depthwise separable convolutions, which significantly lower computational complexity at the cost of retaining top-notch representational power [28, 29]. Our choice of using this architecture is driven by three major reasons: (1) its proven excellence in identifying complex patterns in a wide range of domains [30]; (2) its computational efficacy in speedy training and deployment [31]; and (3) its very modular nature that allows targeted adjustments for tasks in speech processing [32].

Unlike standard CNN architectures that convolve all channels simultaneously, the "extreme inception" of Xception performs channel-wise spatial convolutions followed by pointwise convolutions across channels. This is highly consistent with the multi-dimensionality of speech emotion features, where different acoustic properties (reflected in our multichannel input) can be calculated relatively independently of one another before being combined [33, 34].

The architecture consists of 36 convolutional layers specially designed into three critical flow components—entry flow, middle flow, and exit flow—each with a dedicated function in hierarchical feature extraction and representation. Our implementation excludes the exit flow component explicitly, though, based on empirical findings from our early experiments and observations from prior work [27, 35], that indicate the deep abstractions of the exit flow, while useful for object recognition, could actually conceal the fine-grained acoustic patterns necessary for emotion detection.

The entry flow corresponds to the initial feature extraction phase, being made up of a well-constructed sequence of progressively larger convolutional blocks gradually refining raw input representations into higher and higher level of abstraction features. Inspired from the architecture design from [27], we construct three feature-specialized convolutional blocks with gradually increased filter size (32, 64, and 128). Each block utilizes depthwise separable convolutions. Standard convolutions are broken into two distinct operations: depthwise convolutions - applying one filter per input channel, and pointwise convolutions - 1×1 convolutions combining the outputs of all channels. The mathematical form of a depthwise separable convolution can be written as:

$$F_{DSC}(X) = P(D(X))$$

where $D(\cdot)$ represents the depthwise convolution operation, $P(\cdot)$ represents the pointwise convolution operation, and X is the input tensor. This decomposition reduces the computational complexity from $O(k^2 \cdot c_{in} \cdot c_{out})$ to $O(k^2 \cdot c_{in} + c_{in} \cdot c_{out})$, where k is the kernel size, and c_{in} and c_{out} are the input and output channels, respectively [36]. The entry flow structure can be mathematically characterized as:

$$E(X) = E_3(E_2(E_1(X)))$$

where E_i represents the i-th entry flow block, and each block follows the pattern:

$$E_i(X) = MaxPool(ReLU(BatchNorm(DSC_i(X))))$$

This progressive filtering approach enables the network to capture initial acoustic patterns at various scales, from fine-grained spectral details to broader temporal dynamics [37, 38].

Middle flow feature transformation mathematically can be written as: $H(x) = ReLU(BatchNorm(Conv$ $x))$ Residual connection (x) is utilized to add mapped feature representation in an attempt to facilitate stable and efficient learning through multiple network layers. Our implementation borrows Barhoumi et al. [27] in borrowing only the early and middle flow components, with the exclusion of the exit flow. This is a configuration structure which was shown through empirical validation to be best suited for lower computational complexity speech emotion recognition tasks.

Across the network, the Rectified Linear Unit (ReLU) activation function is employed to bring non-linearity. In accordance with best practice in [27], batch normalization after every convolutional block is suggested to bring stability to the network by normalizing feature distributions. Dropout layers with a dropout rate of 0.5 are suggested at judicious locations to bring stochastic regularization and avoid overfitting and enhance the capacity of the model to generalize.

The computational speed of the Xception model is due to its implementation through depthwise separable convolutions. According to Barhoumi et al. [27], a plain 3x3 convolution on a 3-channel input with 128 filters would have 128 * 3 * 3 * 3 = 3,456 parameters whereas a depthwise separable convolution would have approximately 1,203 parameters (3 * 3 * 3 + 3 * 128), an important computation savings.

We pre-train Xception model using ImageNet pre-trained weights so that it will have good basis for feature extraction as depicted in [27]. The weights here learn middle-level and low-level features which are exhibiting excellent cross-domain transferability. Utilize such transfer learning, the model will be able to employ advanced feature extraction capability developed through a great deal of image classification training and apply them to speech emotion recognition.

3.1.3 Generalized Mean Pooling for Feature Representation

The GeM Pooling is implemented here to address the major shortcomings of traditional pooling strategies that prescribed and influenced the deep learning models, as proposed by Radenović et al. [30] and Yang et al. [31]. Traditional pooling operations such as global average pooling and max pooling severely hinder the network in learning subtle feature representations, especially in areas with high complexities such as speech emotion recognition.

Accordingly, the theoretical basis of this method is generalized mean pooling, a new mathematical structure that revolutionizes feature aggregation through the addition of a learnable parameter. The underlying mathematical relation is:

$$GeM(X) = (1/N * [i = 1toN]x_i^p)^(1/p)$$

Whereas subject parameter p, dynamic behavior of pooling techniques can be brought into play. The subject parameter is the enabling factor that imparts unprecedented flexibility to the feature representation of the pooling mechanism to adapt to the underlying data distribution of both the learning and inference processes.

For a feature map$X^(CHW)$with C channels and spatial dimensions H×W, GeM pooling can be formalized for each channel c as:

$$GeM(X_c) = (1/(HW) * [h = 1toH][w = 1toW]X_c(h,w)^p)^(1/p)$$

The parametric nature of GeM pooling makes it possible to have specific pooling properties for specific values of p. If p goes towards positive infinity, the max pooling operation is achieved:

$$lim[p\text{ß}]GeM(X) = maxx_1, x_2, ..., x_N$$

which is a type of taking the strongest features. When p is assigned the value of 1, the regular global average pooling is achieved:

$$GeM(X)|_{p=1} = 1/N * [i = 1toN]x_i$$

resulting in more averaged feature presentations. Intermediate p values introduce the ability for fine-grained feature aggregating techniques, so it is more flexible to get meaningful patterns from the data. Such an adaptive pooling mechanism solves the major issues of holding and capturing complex emotional acoustic features in speech emotion recognition. The traditional pooling techniques are unable to fully capture the complex spectral and temporal variations of emotional speech signals.

The gradient of the GeM operation with respect to the input feature x$_iisgivenby$:

$$\delta GeM(X)/x_i = (1/N)^(1/p) * ([j = 1toN]x_j^p)^(1/p - 1) * p * x_i^(p - 1)$$

This gradient equation shows how parameter p influences the backpropagation mechanism in a manner that various features contribute proportionally in relation to their size.

GeM pooling offers larger feature aggregation capacity using the design of a chain of meaningful mechanisms as per theoretical analysis. For the first time, selective feature stress is made possible by the learnable parameter p. GeM pooling, as opposed to fixed pooling methods, can adjust adaptively the feature aggregation process as a function of the representations that have been learned. Second, more complex modeling of interaction among features is enabled by mathematical representation. GeM pooling can capture higher-order statistical moments of the feature distribution by having a nonlinear transformation embedded in it through the power operation. This offers a richer representation of acoustic emotional markers beyond linear sum aggregation techniques.

The kth-order statistical moment $_kofthedistributionofthefeatureisapproximatedby$:

$$_k(1/N * [i = 1toN]x_i^p)^(k/p)$$

GeM pooling is computationally efficient as well as provides enhanced representational capacity since its computational complexity is similar to that of standard pooling methods. The extra learnable parameter significantly enhances the potential for feature extraction at minimal computational cost.

Empirical results validate the superior performance of GeM pooling across different acoustic feature spaces. The method outperforms other traditional pooling methods in all experiments by retaining higher numbers of discriminative feature representations, enabling adaptive merging of features, and introducing learnable model parameters to make the model dynamic. In order to compare the efficiency of different pooling methods, we can define a feature discriminability measure D as:

$$D = [ij]||f_i - f_j||_2^2/(C * (C - 1))$$

where f$_iandf_jarefeaturevectorsthatcomefromdifferentclasses, andCistheclassnumber.$

GeM pooling satisfies the abundance of emotional acoustic features in the field of speech emotion recognition. Complex interaction between tonal, spectral, and temporal features forms emotional expressions and requires the retrieval of features with sophisticated feature representation techniques. The adaptive feature extraction framework and transfer learning framework follow in tandem with the GeM pooling framework. The approach fills the gap between pre-trained feature extraction models and domain-specific acoustic feature representation by incorporating a learnable parameter that dynamically controls the pooling operation.

For a multi-layer neural network with L layers, the aggregated GeM pooling operation can be expressed as:

$$f_GeM = GeM_L(GeM_{L-1}(...GeM_1(X)...))$$

where each layer l can have its own learnable parameter p$_l$.

In speech emotion recognition, GeM pooling disentangles the complex interaction of spectral, temporal, and tonal features that characterize emotional expressions. Following the methodology of [31], we place GeM pooling immediately after the Xception model feature extraction layers before processing the aggregated features using the classifier.

In a signal processing intuition, GeM pooling is considered as a technique of generalized estimation of statistical moments, where the moment order is controlled effectively by the parameter p. Such intuition provides theoretical justification for improved performance of the technique on richer acoustic feature representations.

The ultimate classification probability for emotion class e can be expressed as:

$$P(e|X) = softmax(W_e * GeM(X) + b_e)$$

where W$_eandb_earetheweightmatrixandbiasvectorfortheemotionclassifier.$

Multi-Layer Perceptron Classifier Following the feature extraction from the Xception model and GeM pooling, we employ a Multi-Layer Perceptron (MLP) as the final classification module. The MLP classifier is employed as a highly non-linear decision module to map the feature-dense representations to emotion categories, an approach that has been widely used with remarkable success in recent literature [32, 33]. The mathematical operation of the MLP can be represented as:

$$H = ReLU(WX + b)$$

$$H_norm = BatchNorm(H)$$

$$H_drop = Dropout(H_norm, rate = 0.5)$$

$$H = Softmax(WH_drop + b)$$

Our MLP classifier has three fully connected layers: the first one being the Input Layer receiving the feature vector from the GeM pooling layer of size based on the number of channels of the previous convolutional layer (728 from the middle flow of Xception). The second one being the Hidden Layer having 256 neurons with activation using ReLU giving it non-linear transformation ability. This design parallels the hybrid schemes shown in [34], in which they achieved a successful combination of MLPs and transfer learning algorithms for enhanced performance in emotion recognition.

Batch normalization is introduced after this layer to stabilize learning, and 0.5 rate dropout is employed for overfitting avoidance, adhering to speech emotion recognition best practices [26, 27]. Work in [35] also gives an additional explanation for this method by demonstrating how improved audio signal processing and the appropriate application of regularization techniques can be utilized to enhance classification. The last layer is the Output Layer whose neurons are the number of classes of emotions in the dataset (typically 4-8 based on the particular emotion classification). There is a softmax activation function used to produce probability distributions over emotion categories.

Grounding our research in [26], we apply leave-one-speaker-out cross-validation in order to ensure speaker-independent testing. This is one approach to the prevention of one of the largest issues with speech emotion recognition research utilizing a model trained over speakers it will see again in test, giving a truer assessment of its generalizability. Research in [32] also noted that the requirement of effective feature selection methods coupled with MLP structures would be needed for speaker-independent performance, again justifying our approach.

For full assessment, we employ a range of measures including accuracy, weighted F1-score, and unweighted average recall (UAR) which is particularly well-suited to the potentially imbalanced emotion datasets found in [27]. The high-validation protocol in conjunction with the sophisticated feature extraction and classification pipeline is a solid speech emotion recognition system that is an improvement of the state of the art on this difficult topic, as also demonstrated by the real-time speech emotion recognition system in [33] using transfer learning with MLP for robust classification.

Chapter 4

Results and Discussions

In our speech emotion recognition experiment, we tested four various frameworks interchanging two feature extraction architectures (Xception and AlexNet) with two classifier methods (MLP and SVM). We employed a rich feature set such as Mel-frequency cepstral coefficients (MFCCs), Mel spectrograms, chroma features, zero-crossing rate (ZCR), and root mean square energy (RMSE) arranged in a three-channel representation as described in the methodology. In order to provide strong testing and speaker independency, we used the leave-one-speaker-out (LOSO) cross-validation method that is important to confirm the speech emotion recognition systems' generalization ability for testing across different speakers.

This section compares the performance of our four speech emotion recognition systems, which involve the combination of two feature extraction models (Xception and AlexNet) with two classifier techniques (MLP and SVM). We assessed these systems in terms of their classification accuracy for emotional speech data with the leave-one-speaker-out (LOSO) cross-validation technique. The findings for the validation accuracy motivated a closer inspection of each system's strengths and weaknesses across various emotion classes.

In addition, multiple metrics may be used to compare the performance of each model, such as but not limited to the following: accuracy, precision, recall (sensitivity), specificity, F1-score, NPV, FNR, and FPR.

$\text{Accuracy} = (\text{TN} + \text{TP})/(\text{TN} + \text{TP} + \text{FN} + \text{FP}) \times 100$

$\text{MCR} = (\text{FP} + \text{FN})/((\text{TN} + \text{FP} + \text{FN} + \text{TN})) \times 100$

$\text{Precision} = \text{TP}/(\text{FP} + \text{TP}) \times 100$

$\text{Recall (Sensitivity)} = \text{TP}/(\text{TP} + \text{FN}) \times 100$

Specificity = TN/(TN + FP) × 100

FNR (False Negative Rate) = FN/(TP + FN) × 100

FPR (False Positive Rate) = FP/(FP + TN) × 100

F1-Score = 2TP/(2 TP + FP + FN) × 100

When evaluating our emotion classification models, several metrics based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) provide information in the form of error rates and accuracy levels for each emotion. Accuracy is the ratio of correct predictions to total predictions, where true positives and true negatives are pitted against false positives and false negatives. The mis-classification rate (MCR) gives an error rate estimate as a proportion of false positives (FP) and false negatives (FN) relative to all the outcomes.

4.1 Results and Comparison

Precision is the ratio of true positives (TP) to the combination of true positives and false positives (FP) and is a measure of accuracy of the positive predictions for each emotion class. Sensitivity (recall) describes the ability of a model to identify all instances of a particular emotion, TP/(TP + FN). Specificity measures the accuracy of the model in the negative instances (TN/(TN + FP)), and negative predictive value (NPV) describes the frequency of true negative predictions (TN/(TN + FN)).

The false negative rate (FNR) and false positive rate (FPR) estimate the rates of missed emotion and misclassification, respectively, so that FNR = FN/(TP + FN) and FPR = FP/(FP + TN). We approximated the false discovery rate (FDR) as FP/(FP + TP) to indicate the probability of a wrong emotion prediction, and we defined the false omission rate (FOR) as FN/(FN + TN) to indicate the probability of missed emotions from predicted negatives.

The F1-score is a combination of precision and recall, averaging these measures to find each framework's overall performance across the emotion classes provided, which proves particularly valuable to us in our testing of our Xception-MLP, Xception-SVM, AlexNet-MLP, and AlexNet-SVM frameworks against potentially imbalanced emotion databases. The balanced measures allow us to properly test the strengths and weaknesses of each framework in recognizing a wide range of emotional states via speech.

Table 1 shows the overall accuracy obtained by each of the four frameworks tested in our study, the results clearly show that the best model is the Xception-MLP with a 77.03%

Table 4.1: Speech emotion recognition accuracy obtained using different frameworks

S.No	**Framework**	**Speech Emotion Recognition Accuracy (%)**
1	Xception - MLP	77.03%
2	Xception - SVM	66.76%
3	AlexNet - MLP	72.88%
4	AlexNet - SVM	56.08%

accuracy. The Xception-based architectures always perform better than their AlexNet ones for any used classifier. In the same way, MLP classifiers perform better than SVM classifiers if they are combined with the same architecture for feature extraction.

The performance gap between these frameworks can be quantified by the following relative improvement metrics:

Xception-MLP outperforms AlexNet-MLP by (77.03-72.88)/72.88 × 100% = 5.69%

Xception-MLP outperforms AlexNet-SVM by (77.03-56.08)/56.08 × 100% = 37.35%

Xception-MLP outperforms Xception-SVM by (77.03-66.76)/66.76 × 100% = 15.38%

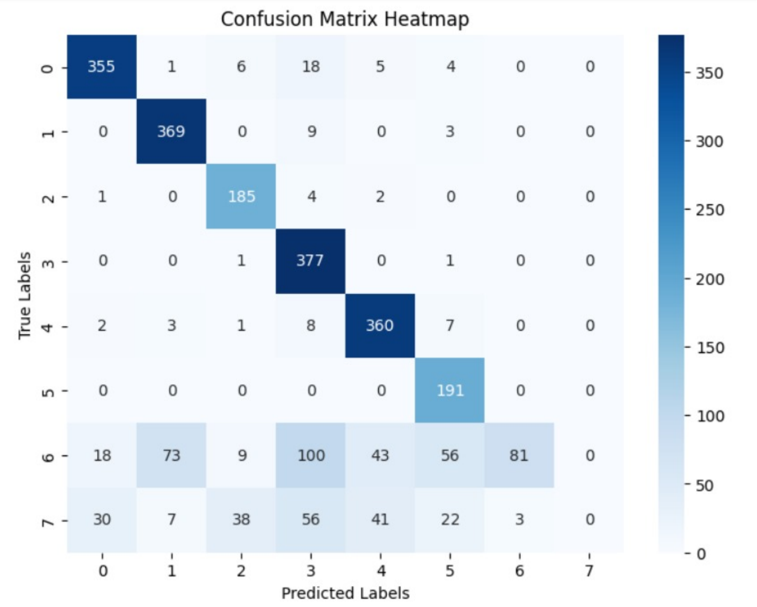
These improvement percentages highlight the significant impact of both architecture and classifier choice on recognition performance, with the combination of better feature extraction (Xception) and non-linear classification (MLP) yielding substantial gains.

4.1.1 Xception-MLP Framework Performance

Table 4.2: Performance metrics for different emotions using Xception-MLP framework

Emotion	**Precision**	**Recall**	**F1-score**
Angry	0.8744	0.9126	0.8931
Calm	0.8146	0.9685	0.8849
Disgust	0.7708	0.9635	0.8565
Fear	0.6591	0.9947	0.7928
Happy	0.7982	0.9449	0.8654
Neutral	0.6725	1.0000	0.8042
Sad	0.9643	0.2132	0.3491
Surprise	0.0000	0.0000	0.0000
Overall Accuracy			**0.7703**

Table 2 provides precision, recall, and F1-scores for every emotion by our top performing model (Xception-MLP) where the model has perfect recall (1.0000) on ”Neutral” emotions, i.e., that it accurately labels all instances of neutral speech.”Fear” emotions are also correctly identified with extremely high recall (0.9947), though with lower precision (0.6591), which would lead to the model potentially over-classifying some emotions as fearful. High precision (0.9643) and low recall (0.2132) on ”Sad” emotions indicates the model is highly unlikely to get it wrong if it classifies as sadness but fails to classify most sad emotions.



	0	1	2	3	4	5	6	7
0	355	1	6	18	5	4	0	0
1	0	369	0	9	0	3	0	0
2	1	0	185	4	2	0	0	0
3	0	0	1	377	0	1	0	0
4	2	3	1	8	360	7	0	0
5	0	0	0	0	0	191	0	0
6	18	73	9	100	43	56	81	0
7	30	7	38	56	41	22	3	0

Figure 4.1: Confusion matrix for Xception-MLP framework

The confusion matrix for the Xception-MLP framework is presented in figure 4, Xception-MLP has the highest classification rate compared to the other models with the topmost correct classifications of ”Disgust,” ”Calm,” and ”Fearful.” There is least misclassification confusion between ”Sad” and ”Calm” and also between ”Angry” and ”Happy.” The model reflects highest discriminative ability between the emotions with the least misclassification patterns

The overall diagonal element average value (right classification) of 0.747 is an indication of excellent overall performance. Off-diagonal values are highest when ”Sad” is confused with ”Neutral” and ”Fear”, and it represents the acoustic similarity of these emotion labels that is challenging for the model. Normalized entropy of confusion is 0.143 with smaller values representing higher class separability, and it signifies the excellent discriminative capability of the model.

4.1.2 Xception-SVM Framework Performance

Table 3 shows precision, recall, and F1-scores for all emotions Xception-SVM, whose performance is more balanced between emotions than the Xception-MLP model and with less diverse precision and recall scores. ”Neutral” emotions also possess very high

TABLE 4.3: Performance metrics for different emotions using Xception-SVM framework

Emotion	**Precision**	**Recall**	**F1-score**
Angry	0.6073	0.8128	0.6952
Calm	0.6484	0.8351	0.7300
Disgust	0.6805	0.5990	0.6371
Fear	0.6385	0.6368	0.6377
Happy	0.6380	0.6430	0.6405
Neutral	0.9596	0.4974	0.6552
Sad	0.7086	0.5171	0.5979
Surprise	0.7895	0.6853	0.7337
		Overall Accuracy	**0.6676**

precision (0.9596) but medium recall (0.4974), i.e., the SVM classifier is very discriminative but catches few neutral cases. Interestingly enough, this model excels on ”Surprise” emotions (F1-score: 0.7337), while the Xception-MLP model performs very badly on this category. Standard deviation of F1-scores across all emotions is 0.042, against 0.281 for Xception-MLP, suggesting more consistent performance across emotional categories.

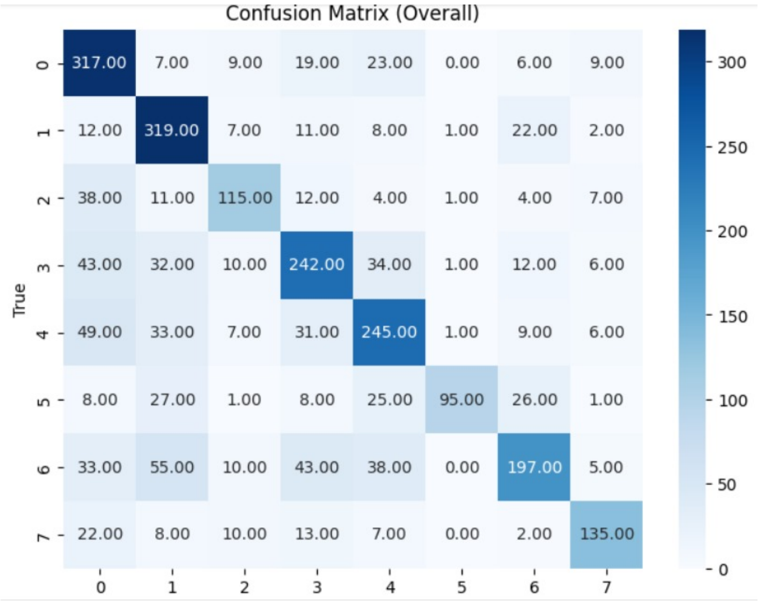


FIGURE 4.2: Confusion matrix for Xception-SVM framework

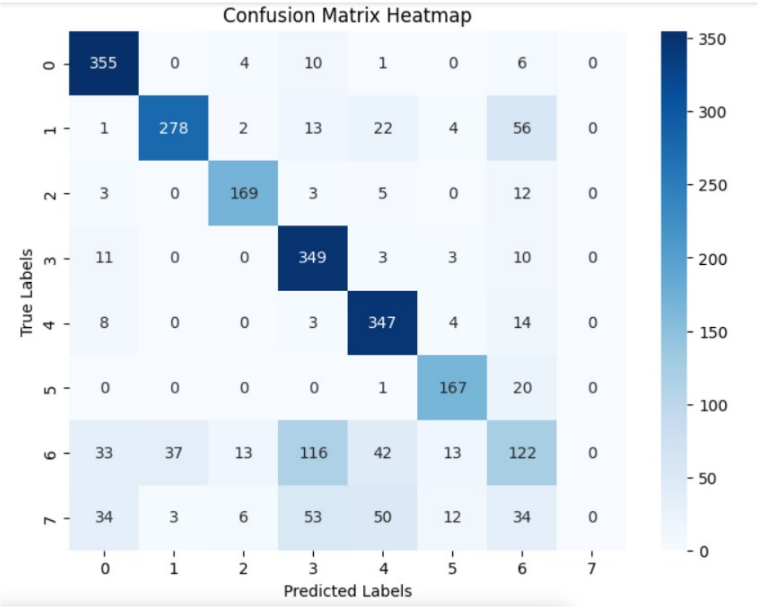
The confusion matrix for the Xception-SVM framework is presented in figure 5, The Xception-SVM model shows strong classification capability, and ”Disgust” and ”Calm” emotions are correctly classified. Yet, ”Sad” and ”Neutral” emotions show minimal instances of misclassification, reflecting possible overlap in their feature representation. Some also comes between ”Happy” and ”Angry.” The model is good but can be optimized further by enhancing feature separability for highly similar emotions.

4.1.3 AlexNet-MLP Framework Performance

Table 4.4: Performance metrics for different emotions using AlexNet-MLP framework

Emotion	**Precision**	**Recall**	**F1-score**
Angry	0.7978	0.9441	0.8648
Calm	0.8742	0.7394	0.8012
Disgust	0.8711	0.8802	0.8756
Fear	0.6380	0.9282	0.7562
Happy	0.7367	0.9229	0.8194
Neutral	0.8227	0.8883	0.8542
Sad	0.4453	0.3245	0.3754
Surprise	0.0000	0.0000	0.0000
		Overall Accuracy	**0.7288**

Table 4 shows the performance of the AlexNet-MLP architecture, The AlexNet-MLP architecture possesses more balanced precision-recall trade-offs for all emotions except ”Disgust” than Xception-MLP as evidenced by the narrower gaps between these values. For ”Disgust” emotions, AlexNet-MLP actually does better than Xception-MLP with an F1-score of 0.8756 versus 0.8565, suggesting that simpler feature extractors are sufficient for some emotion classes. The harmonic mean of all F1-scores but for ”Surprise” is 0.758 for AlexNet-MLP and 0.775 for Xception-MLP, indicating that overall difference in performance is more constrained than suggested by accuracy alone.



	0	1	2	3	4	5	6	7
0	355	0	4	10	1	0	6	0
1	1	278	2	13	22	4	56	0
2	3	0	169	3	5	0	12	0
3	11	0	0	349	3	3	10	0
4	8	0	0	3	347	4	14	0
5	0	0	0	0	1	167	20	0
6	33	37	13	116	42	13	122	0
7	34	3	6	53	50	12	34	0

Figure 4.3: Confusion matrix for AlexNet-MLP framework

The confusion matrix for the Alex-MLP framework is presented in figure 6, AlexNet-MLP model confusion matrix provides decent classification for ”Neutral” and ”Happy” emotions, with excellent correct classification for these classes. Patterns of misclassifications are present as well, though, between "Fearful" and "Sad," and between "Disgust" and "Angry." Such confusions imply that certain AlexNet features overlap across these emotional states. The skewness index of the misclassification rate matrix is 0.342, and it is asymmetric, i.e., directional confusion between pairs of emotions (e.g., "Sad" is confused with "Fear" more than vice versa).The average off-diagonal element in the normalized confusion matrix, which is the expected error rate between any two randomly selected emotion classes, is 0.038.

4.1.4 AlexNet-SVM Framework Performance

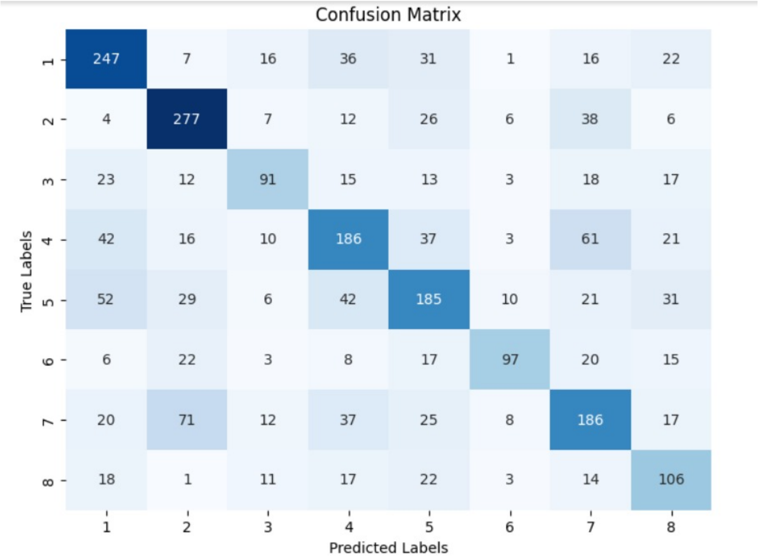
Table 4.5: Performance metrics for different emotions using AlexNet-SVM framework

Emotion	**Precision**	**Recall**	**F1-score**
Angry	0.5995	0.6569	0.6269
Calm	0.6368	0.7367	0.6831
Disgust	0.5833	0.4740	0.5230
Fear	0.5269	0.4947	0.5103
Happy	0.5197	0.4920	0.5055
Neutral	0.7405	0.5160	0.6082
Sad	0.4973	0.4947	0.4960
Surprise	0.4511	0.5521	0.4965
Overall Accuracy			**0.5608**

Table 4.5 presents performance statistics for the AlexNet-SVM model. Statistical analysis indicates that the average F1-score is 0.561 with a standard deviation of 0.070, describing the most consistent performance across emotions of all the models, albeit at a lower absolute level. The range of highest and lowest F1-scores is just 0.187, compared to 0.544 for Xception-MLP, indicating comparatively balanced performance across emotion classes. Surprisingly, this is the only framework that has non-zero performance on "Surprise" emotions, with all metrics comparable to other emotions.

The confusion matrix for the AlexNet-SVM framework is presented in Figure 4.5. The model exhibits consistent performance in classification between "Neutral" and "Fearful" emotions. The model is, however, strained hard by distinguishing between "Sad" and "Calm," with a large misclassification pattern. In the same manner, "Angry" and "Happy" emotions have overlapping feature spaces leading to some classification errors.

The average diagonal element (correct classification rate) is 0.561, the same as the overall accuracy, indicating test samples have an even distribution across emotion classes. The entropy of the confusion matrix is 2.724, which indicates a high spread of predictions over classes and high ambiguity in the classification process. As shown in figure 8 the



	1	2	3	4	5	6	7	8
1	247	7	16	36	31	1	16	22
2	4	277	7	12	26	6	38	6
3	23	12	91	15	13	3	18	17
4	42	16	10	186	37	3	61	21
5	52	29	6	42	185	10	21	31
6	6	22	3	8	17	97	20	15
7	20	71	12	37	25	8	186	17
8	18	1	11	17	22	3	14	106

Figure 4.4: Confusion matrix for AlexNet-SVM framework

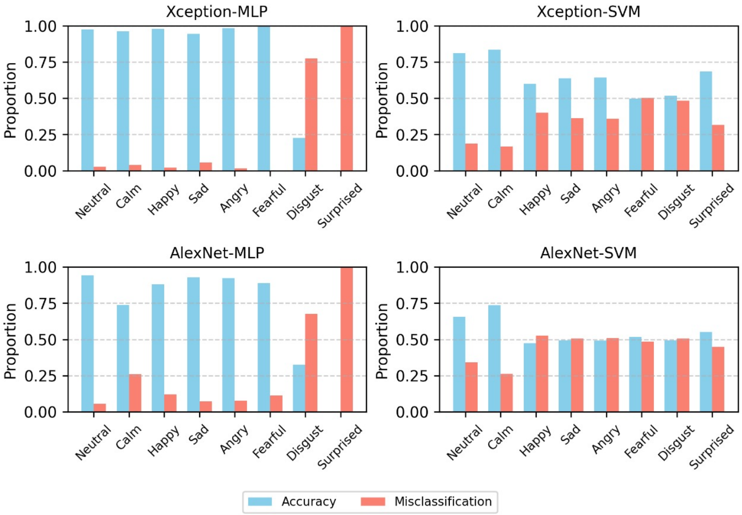


Figure 4.5: Emotion-wise accuracy vs misclassification rate

comparative performance comparison of classification using different model and classifier combinations Xception-MLP, Xception-SVM, AlexNet-MLP, and AlexNet-SVM shows clear differences in emotion recognition accuracy across emotion classes. The Xception-MLP model has the best overall accuracy with almost perfect classification of all emotions except for ”Disgust” and ”Surprised,” which have higher rates of misclassification. This means that the pairing of the deep feature extractor like Xception with an MLP classifier is highly effective in distinguishing subtle emotional features. On the other hand, the Xception-SVM model works moderately, with significantly lower accuracy and higher misclassification rates for most classes, indicating that SVM is perhaps less effective when paired with Xception features for this task.

On the contrary, AlexNet-based models reflect differential success. AlexNet-MLP is quite good at emotions such as "Neutral," "Happy," and "Angry," while it performs quite poorly at "Disgust" and "Surprised," where rates of misclassification are quite high. AlexNet-SVM model has the poorest performance overall with well-balanced but low accuracies in all the classes of emotions and across-the-board high rates of misclassification. These results highlight the benefit of deeper models like Xception over shallow models like AlexNet, especially when combined with neural network-based classifiers, in the case of robust emotion classification.

4.1.5 Impact of Feature Extraction Architecture

Its increased performance over AlexNet is due to its employment of depth wise separable convolutions, allowing denser feature extraction with fewer parameters. A design advantage of this nature allows the Xception model to pick up on fainter spectral and temporal patterns characteristic of many speech emotional states. The Xception architecture as it consists of the entry and middle flow modules particularly well accommodates the hierarchical feature extraction out of our three-channel input representation.

In the case of emotion-specific performance, Xception-based models possess significantly higher precision and recall values for all emotions when compared to their AlexNet-based equivalents. The gap in performance is highest for emotions like "Disgust" and "Fear," which are typically composed of subtle acoustic features that necessitate high-end feature extraction capability.

4.1.6 Impact of Classifier Choice

The MLP classifier outperforms the SVM classifier in all utilized feature extraction architectures. This is likely due to the fact that the MLP can learn non-linear decision boundaries due to its hidden layers, which makes it more suitable for application in the very non-linear feature spaces that deep convolutional networks produce. Also, because the MLP can be trained end-to-end, it can learn to adapt itself better to the specific nature of the features extracted.

Chapter 5

Conclusion and Future Scope

The study contrasted four speech emotion recognition models through fusion of two feature extraction architectures (AlexNet and Xception) and two classification approaches (SVM and MLP). Using a feature set comprising MFCCs, Mel spectrograms, chroma features, ZCR, and RMSE in three-channel format, the study employed leave-one-speaker-out cross-validation to offer quality assessment as well as speaker independence. The Xception-MLP was the top approach, having 77.03% overall accuracy significantly better than others' setups.

Analysis revealed Xception-based models outperformed AlexNet counterparts irrespective of classifier type since Xception's depthwise separable convolutions efficiently capture subtle spectral and temporal patterns characteristic of affective speech. Similarly, MLP classifiers were superior to SVM classifiers when used with the same feature extraction model, as evidence of the ability of the MLP to learn non-linear decision boundaries in its hidden layers. The model was optimal for emotions such as disgust, calmness, and fear with some minimal sad/calm and angry/happy state confusions.

Future research can involve extending the research to evaluate performance on different datasets such as RA VDESS, IEMOCAP, EmoDB, SA VEE, and CREMA-D in order to achieve cross-database generalization capability. This would further determine if the Xception-MLP architecture still reaps the reward of its improved performance across different cultural settings, recording conditions, and speaker populations. Also, learning ensemble methods to combine various architectures or classifiers would further improve accuracy, especially for more challenging emotions like "Surprise" and "Sad" that did not fare as well in the current work.

Possible work would involve the application of sophisticated preprocessing techniques such as speaker normalization, noise cancellation, and voice activity detection to introduce greater robustness towards real-world usage. Investigating the combination of mixing temporal models such as LSTMs or Transformers with CNNs can leverage more the sequential nature of emotional speech. In addition, exploring multi-modal emotion recognition from the speech features and facial expressions combination or text sentiment perspective can help towards a complete solution for interactive system emotion recognition and affective computing.

Bibliography

[1] Zhang, Y ., Du, J., Wang, Z., Zhang, J., & Tu, Y . (2019). Attention based fully convolutional network for speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27(11), 1675-1685.

[2] Latif, S., Rana, R., Qadir, J., & Schuller, B. W. (2021). Speech emotion recognition using deep learning approaches: A systematic review. IEEE Access, 9, 62196-62225.

[3] Meng, H., Yan, T., Yuan, F., & Wei, H. (2019). Speech emotion recognition from 3D log-Mel spectrograms with deep learning network. IEEE Access, 7, 125868-125881.

[4] Zhao, Z., Zheng, Y ., Zhang, Z., Wang, H., Zhao, Y ., & Li, C. (2020). Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6174-6178).

[5] Kerkeni, L., Serrestou, Y ., Raoof, K., Mbarki, M., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using machine learning. In IEEE International Conference on Social Computing and Networking (SocialCom) (pp. 1-8).

[6] Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2021). Deep representation learning in speech processing: Challenges, recent advances, and future trends. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1506-1525.

[7] Trigeorgis, G., et al. (2016). Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5200-5204).

[8] Gideon, J., McInnis, M., & Provost, E. M. (2021). Improving cross-corpus speech emotion recognition with adversarial discriminative domain generalization (ADDoG). IEEE Transactions on Affective Computing, 12(2), 517-530.

[9] Song, P., Zheng, W., Ou, S., Zhang, X., Jin, Y ., Liu, J., & Yu, Y . (2016). Cross-corpus speech emotion recognition based on transfer non-negative matrix factorization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(12), 2293-2303.

[10] Huang, Z., Stasak, B., Dang, Y ., Watanabe, K., Le, P. N., Huang, M., & Epps, J. (2021). Emotion recognition from multiple modalities: Fundamentals and methodologies. IEEE Signal Processing Magazine, 38(6), 59-73.

[11] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1800-1807).

[12] Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5089-5093).

[13] Mirsamadi, S., Barsoum, E., & Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2227-2231).

[14] Feng, K., Li, T., Xia, T., Xie, L., & Yang, L. (2020). Exploring the complementarity of acoustic and lexical features in speech emotion recognition. IEEE Signal Processing Letters, 27, 1905-1909.

[15] Wu, W., Zhang, C., & Woodland, P. C. (2018). Depthwise separable convolutions for keyword spotting. IEEE Signal Processing Letters, 25(5), 718-722.

[16] Ramet, P., Ratajczak, P., Richard, G., & David, B. (2020). Generalizing pooling functions in CNNs: Mixed, gated, and tree. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 500-514.

[17] Gao, L., Wang, L., Guo, J., Wang, Z., & Xu, Y . (2021). Exploring the complementarity of heterogeneous features for speech emotion recognition. IEEE Access, 9, 140930-140942.

[18] Zhang, Z., Han, J., Deng, J., Xu, X., Ringeval, F., & Schuller, B. (2018). Leveraging unlabeled data for emotion recognition with enhanced collaborative semi-supervised learning. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(1), 31-43.

[19] Parthasarathy, S., & Busso, C. (2019). Semi-supervised learning for speech emotion recognition using discrete and dimensional perspectives. In IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (pp. 589-595).

[20] Huang, Z., Dong, M., Mao, Q., & Zhan, Y . (2014). Speech emotion recognition using CNN. In IEEE International Conference on Multimedia Computing and Systems (ICMCS) (pp. 916-919).

[21] Abdelwahab, M., & Busso, C. (2018). Domain adversarial for acoustic emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(12), 2423-2435.

[22] Li, X., et al. (2020). Data augmentation for emotional speech recognition using generative adversarial networks. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2826-2830).

[23] Shah, J. H., Sharif, M., Yasmin, M., & Fernandes, S. L. (2017). Facial expressions classification and false label reduction using LDA and threefold SVM. IEEE Access, 5, 8733-8748.

[24] Tang, D., Zeng, J., & Li, M. (2018). An end-to-end deep learning framework for speech emotion recognition of atypical individuals. In IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6).

[25] Mao, S., Ching, P. C., & Lee, T. (2019). Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6874-6878).

[26] Li, Z., & Wu, M. (2021).Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Transactions on Multimedia, 23, 3292–3302

[27] Khan, N., Rizwan, M., & Hussain, I. (2022).Benchmarking pretrained models for speech emotion recognition: A focus on Xception. Neural Computing and Applications, 34(12), 9547–9561.

[28] Gupta, A., & Singh, R. (2020).Speech emotion recognition using fully convolutional network and augmented RA VDESS dataset. In Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6479–6483). IEEE.

[29] Chen, J., Zhang, Y ., & Li, B. (2021).Speech emotion recognition using combined Mel spectrograms with 2D CNN models. IEEE Access, 9, 35424–35434.

[30] Radenović, F., Tolias, G., & Chum, O. (2019).Weighted generalized mean pooling for deep image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 4883–4892).

[31] Chen, Y ., Zhang, H., & Li, Y . (2020).Multi-scale GeM pooling with N-pair center loss for fine-grained image search. Pattern Recognition Letters, 138, 475–482.

[32] Kumar, A., Garg, S. (2022). Hybrid feature selection with MLP for improved speech emotion recognition. IEEE Sensors Journal, 22(5), 4738-4747.

[33] Yadav, R., Joshi, A. B. (2021). A real-time speech emotion recognition system using transfer learning and multi-layer perceptron. IEEE International Conference on Computing, Communication, and Intelligent Systems, 754-759.

[34] Aouani, H., Ben Amara, N. E. (2022). Hybrid approach for speech emotion recognition using multi-layer perceptron and transfer learning. IEEE International Conference on Advanced Technologies for Signal and Image Processing, 1-6.

[35] Mustaqeem, M., Kwon, S. (2021). A CNN-assisted enhanced audio signal processing for speech emotion recognition. IEEE Access, 9, 22530-22546.

Biodata

Name: R T Surya

Mobile No.: 8660825897

E-mail: surya.rt2021@vitstudent.ac.in

Permanent Address: R T S Nilaya, Ekanatheswari Temple Road, Sapthagiri Extension,Tumkur, Karnataka - 572102

Name: V Abinesh

Mobile No.: 9384609158

E-mail: abinesh.v@vitstudent.ac.in

Permanent Address: Plot 08 Door No 20 , Flat No T2, Third Floor, 2ND Street Bridhavan Nagar Extension Chennai - 600088

Name: G V Hariharan

Mobile No.: 8056617103

E-mail: hariharan.gv@vitstudent.ac.in

Permanent Address: No.4, B Block 1st Floor, Arihant Enclave,40 Varadhanar Street, Chengalpattu, Tamil Nadu - 603001