Emotion Recognition From Human Speeches: Introducing the Three-Stage Pipeline

MD Rishat Sheakh(20301305)
BRAC University
Dhaka, Dhaka, Bangladesh
md.rishat.sheakh@g.bracu.ac.bd

Sartaj Emon Prattoy(20301326)
BRAC University
Dhaka, Dhaka, Bangladesh
sartaj.emon.prattoy@g.bracu.ac.bd

Awon Bin Kamrul(20301367)
BRAC University
Dhaka, Dhaka, Bangladesh
awon.bin.kamrul@g.bracu.ac.bd

KEYWORDS

Three-Stage Pipeline, Transfer Learning, Mel Spectrogram, ResNet18, Inception-v3, GoogLeNet

ACM Reference Format:
MD Rishat Sheakh(20301305), Sartaj Emon Prattoy(20301326), and Awon Bin Kamrul(20301367). 2024. Emotion Recognition From Human Speeches: Introducing the Three-Stage Pipeline. In . ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 ABSTRACT

Speech emotion recognition is a technology to automatically obtain emotion types from given attributive segments. With the increasing demand for emotion recognition in business, education and other fields, the development of high-accuracy speech emotion recognition systems has become a hot research direction in the speech field. Speech emotion recognition takes speech as the carrier of emotion to study the formation and change of various emotions in speech so that the computer can analyze the speaker's specific emotional situation through speech, to make human-computer interaction more humanized. In our paper, we proposed a three-stage pipeline for emotion recognition from speech which includes feature extraction by data augmentation of speech signals and extraction of Mel spectrograms, followed by the use of three pre-trained transfer learning CNN models namely, ResNet18, Inception_v3, and GoogleNet whose prediction scores are fed to the third stage. For the dataset, we will be using the Surrey Audio-Visual Expressed Emotion (SAVEE), the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), and the Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) datasets to evaluate our three-stage pipeline model.

2 INTRODUCTION

Speech emotion recognition (SER) has been increasingly popular among researchers for several decades and shows promise in various fields such as Human-Computer Interaction, Multimedia, and Biomedicine. Speech is a significant means of communication therefore a detailed analysis of speech signals is necessary. Emotion, apart from verbal content, is crucial in speech signals as it can significantly alter the interpretation of a sentence. SER holds potential in speech-enabled interfaces like AI voice assistants, which can monitor emotions to predict psychological changes or signs of mental stress and depression. Its applications extend to medical fields for detecting conditions like Autism and Parkinson's Disease, aiding in educational software for student mental health detection, and enhancing safety in automated vehicles by assessing driver speech for signs of impairment.

SER has a wide range of potential applications, such as helping AI voice assistants monitor emotions for predicting psychological shifts, aiding in medical diagnoses like detecting Autism and Parkinson's Disease, identifying mental health concerns in educational environments, and improving safety in automated vehicles by evaluating the emotional state of drivers. The process of SER involves two essential phases: feature extraction and classification. Feature extraction includes temporal features like signal energy and spectral features obtained through Fourier Transforms. Deep learning, particularly 2D convolutional neural networks (CNN), has been successful in image classification tasks, leading to the proposal of using Mel Spectrogram features for SER.

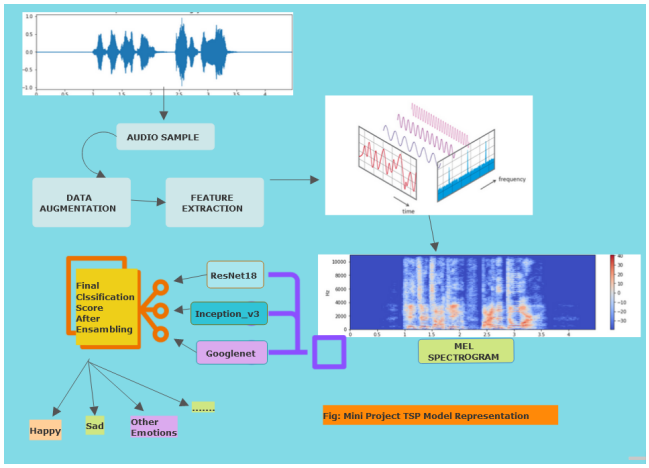

Figure 1: Three Stage Pipeline(TSP) model representation.

3 MOTIVATION AND CONTRIBUTION

As modern technology is evolving day by day with all the automatons with gestures and voice, Speech Emotion Recognition (SER) plays a crucial role in the development of cutting-edge Human-Computer Interaction interfaces. From voice-enabled security devices and authentication systems to Automated Vehicle Environments, emotion can be analyzed to prevent identity mismatches or accidents. Many different fields can benefit from the classification of emotion from its audience, for example, the medical sciences field too can benefit from the classification of emotion from patients' speech for treating Parkinson's disease and Autism to name a few.

Conference'24, May 2024, Brac University, Dhaka, Bangladesh

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3.1 Key points of our proposed model:

Building an end-to-end SER model requires a substantial amount of data. Due to limited labeled audio data availability, we opted to incorporate three pre-trained transfer learning models—Inception-V3, GoogLeNet, and ResNet18—into our ensemble pipeline. Training audio samples on a single classification model can lead to imbalance issues. The Ensemble approach gives an aggregate opinion to all the individual models thereby decreasing noise and giving better and unbiased prediction scores. This makes it a novel approach for SER.

Comparative analysis demonstrates that our Three-Stage-model model outperforms other contemporary SER approaches, validating the novelty and effectiveness of our framework.

The Three-Stage-model model is trained and evaluated on three open-source datasets—SAVEE, CREMA-D, and RAVDESS— to achieve state-of-the-art accuracy compared to current methodologies.

4 LITERATURE REVIEW

Abdul-hadi et al. (2020) [1] conducted a study on emotion recognition techniques using SVM classification based on facial expressions and speech signals. The goal was to Enhance emotion recognition in human-computer interaction systems by integrating information from both facial expressions and speech signals effectively. He used preprocessed video data, extracting features using histograms of oriented gradients (HOG) from facial images and speech signals, and training the SVM model to get the outcome with an accuracy was 85.72% of e the emotion of crying and laughing by speaking.

Singh et al. (2021) [9] proposed a hierarchical deep learning approach for SER, integrating acoustic features and hierarchical DNNs to enhance emotion classification accuracy. The framework utilizes 33 acoustic features extracted from audio data, enabling comprehensive representation. By employing hierarchical DNNs, emotions are processed and classified hierarchically, potentially capturing intricate patterns. The superiority of this approach is evidenced by its outperformance of recent SER techniques. Experimental results demonstrate promising accuracies, with reported rates of 81.2%, 81.7%, and 74.5% on SAVEE, RAVDESS, and IEMOCAP datasets, respectively, showcasing the framework's robustness and applicability across domains.

The study [8] on "Automatic Speech Emotion Recognition Using Support Vector Machine" highlights the importance of SER in Human-Computer Interaction. By analyzing the Berlin Emotional Speech Database, the research extracts features like energy, pitch, LPCC, MFCC, and LPCMCC, employing SVM-based classification. Combining these features significantly boosts recognition accuracy, with the best-achieved accuracy of 82.5% using energy, pitch, and LPCMCC. Future work aims to improve feature extraction techniques and explore real-time applicability for enhanced practical use of SER systems.

Luna-Jiménez et al. (2021) investigate the efficacy of multimodal emotion recognition systems by integrating speech and facial information [6]. Methodologies involve employing transfer learning techniques, fine-tuning CNN-14 for speech emotion recognition, and utilizing STN for facial emotion recognition. The study achieves an accuracy of 80.08% on the RAVDESS dataset. However, limitations are noted, including potential challenges in generalizability across diverse demographic groups and conditions, emphasizing the necessity for further research to address these issues.

This paper [3] tackles the challenge of speech emotion recognition, acknowledging its significance in applications like audio surveillance and clinical studies, while emphasizing existing limitations in feature selection and noise interference. Introducing a new method, the paper extracts five diverse features from sound files and employs a 1-D Convolutional Neural Network (CNN) model to enhance generalization and classification accuracy. Methodologically, features are stacked into a one-dimensional array and fed into the CNN model, with incremental adjustments made for improved accuracy. While achieving state-of-the-art results, the study acknowledges limitations concerning dataset specificity and potential performance influences from factors like dataset size and noise levels. Despite these constraints, the proposed model shows promise for diverse applications, indicating the need for further research to ensure its robustness across various datasets and real-world scenarios in fields like Artificial Intelligence and Mobile Health.

This paper addresses the crucial need for accurate speech-emotion recognition to facilitate effective human-robot interaction across diverse linguistic and cultural contexts. By proposing an ensemble learning approach, the study aims to improve emotion detection accuracy within and across different language corpora, utilizing spectral and prosodic features extracted from audio files. Through experiments conducted on four corpora in English, Urdu, German, and Italian, the proposed method demonstrates significant enhancements in emotion classification performance, particularly in cross-corpus scenarios. However, the study is limited by its focus on a limited set of languages and corpora, with potential implications for generalizability. Nonetheless, the findings underscore the potential of ensemble learning in advancing speech emotion recognition systems for real-world applications, warranting further investigation into scalability and robustness (Zehra et al., 2021).

Lieskovská et al. proposed a paper [4] explores the significance of emotions in human-computer interaction (HCI), emphasizing their integration into various applications such as customer service and mental health care through accurate emotion recognition, particularly in speech signals. By reviewing recent advancements in speech emotion recognition (SER) with a focus on attention mechanisms in deep learning-based solutions, the study contributes to understanding how neural architectures can better capture emotional cues in speech. Through a systematic literature review and comparison of SER systems' accuracy on benchmark databases, the paper underscores the importance of attention mechanisms in improving SER performance while acknowledging potential limitations in coverage and accessibility. Overall, it provides valuable insights into the role of attention mechanisms in HCI and highlights avenues for future research in emotion recognition technology

Another paper proposed by Wang et al. (2020) [10] explores advancements in Speech Emotion Recognition (SER) within Automatic Speech Recognition (ASR) using deep learning techniques, aiming to enhance human-computer interaction. It introduces a dual-level model for processing MFCC features and mel-spectrograms separately, alongside a novel Dual-Sequence LSTM (DS-LSTM) architecture and a data preprocessing mechanism using nearest-neighbor interpolation. Utilizing the IEMOCAP dataset, the study demonstrates the effectiveness of these approaches in improving SER

accuracy. However, reliance on a single dataset may limit generalizability, and the scalability and computational costs of the proposed preprocessing method need further exploration. Overall, the research highlights the potential of ASR and SER to transform human-computer interaction.

This paper(9054629) explores advancements in Speech Emotion Recognition (SER) within Automatic Speech Recognition (ASR) using deep learning techniques, aiming to enhance human-computer interaction. It introduces a dual-level model for processing MFCC features and mel-spectrograms separately, alongside a novel Dual-Sequence LSTM (DS-LSTM) architecture and a data preprocessing mechanism using nearest-neighbor interpolation. Utilizing the IEMOCAP dataset, the study demonstrates the effectiveness of these approaches in improving SER accuracy. However, reliance on a single dataset may limit generalizability, and scalability and computational costs of the proposed preprocessing method need further exploration. Overall, the research highlights the potential of ASR and SER to transform human-computer interaction (Wang et al., 2020).

Mustaqeem et al. (2020) [7] introduce a novel deep learning-based approach for speech emotion recognition (SER), aiming to address challenges in existing methods by combining RBF-based K-means clustering and deep BiLSTM networks. Motivated by the importance of accurately identifying emotions from speech signals, the proposed framework segments audio files selects key segments using RBF-based similarity measurement, and converts them into spectrograms for feature extraction. Leveraging the ResNet101 model, discriminative features are extracted, normalized, and fed into deep BiLSTM networks to learn temporal dependencies and recognize emotional states. While promising results on benchmark datasets demonstrate the potential of the approach, limitations include reliance on specific datasets and computational complexity, suggesting avenues for future research to enhance scalability and robustness across diverse applications.

5 DATASET

Speech Emotion Recognition is a classification problem. The easiest sentiment analysis definition for speech recognition is that it is a classification task. Therefore, it can only be repeated with the quality and quantity of correctly labeled data. Emotion perception is subjective as different people can hear a sense of the same speech as different ones. Consequently, when we choose a dataset, interest is taken in those labeled by people of the same demographic as the base creators. We use simulated databases to mitigate the ways made dialect, allowing some actors and speakers to reproduce the expected feelings by performing prepared lines. In our Research, we have utilized the public SAVEE[] and RAVDESS [5] datasets. These datasets are free of noise and have correctly tagged audio samples. The datasets comprise thorough tables on which data is tagged and how audio samples are particularly distributed across several emotional streams. The detailed description of these datasets is as follows:

5.1 RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song(RAVDESS) [5] contains 1440 files. There are 60 trials per actor and it has 24 actors. 12 of them are male and 12 of them are female. So 1440 files vocalize two lexically matched statements in a neutral North American accent. Speech emotions include calm, happy, sad, angry, fearful, surprised, and disgusted expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. The ratings were provided by 247 individuals for reliability assessment.

5.2 SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) database was recorded from 4 native male English speakers. The text material consisted of 15 TIMIT sentences per emotion. These include 3 common, 2 emotion-specific and 10 generic sentences that were different for each emotion and they were phonetically-balanced. There were seven emotions that were anger, fear, disgust, surprise, sad, happy, and neutral. The file in the dataset includes 480 audio samples in .wav format.

5.3 CREMA-D

Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) is a Diverse Dataset that includes 7,442 original clips from 91 actors. Varied demographics: 48 male, 43 female, spanning ages 20-74, multiple races/ethnicities.Emotion Representation: 12 sentences expressing 6 emotions at 4 intensity levels. Emotions include Anger, Disgust, Fear, Happy, Neutral, Sad.

6 METHODOLOGY

In this project, we followed a certain framework using the datasets mentioned in the previous section. In the first stage, we began with data preparation. The later segments include data augmentation, followed by feature extraction. The extracted data then goes through our selected pre-trained model through a transferred learning process model.

6.1 DATA PREPARATION

The two different datasets that we are working with require a data frame for storing all the emotions of the data in a data frame with their paths. The data frame is used in feature extraction for model training. The Data from all these datasets did go through Visualization and Exploration where each emotion count was plotted and wave plots and spectrograms for audio signals were shown.

6.2 DATA AUGMENTATION

Data augmentation is the process by which we create new synthetic data samples by adding small perturbations to our initial training set. For Audio data we can apply noise injection, shifting time, changing pitch and speed to generate syntactic data. Various disproportionate data can cause problems in the prediction. Thus, to overcome these problems we have to use augmentation on the datasets that we are using. From the various augmentation techniques mentioned above, we are using shifting time with the equation:

$$X_{\text{new}} = X[n \pm s] \tag{1}$$

Here, Xn is the audio signal and s is the number of samples shifted.

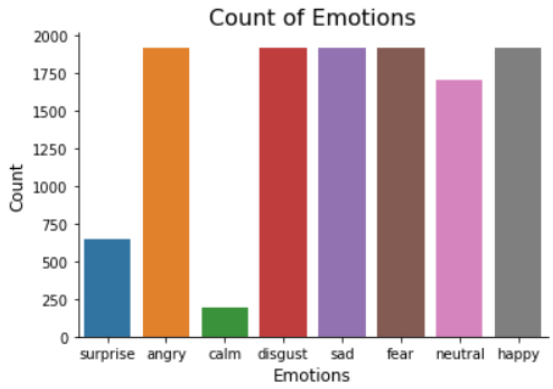

Figure 2: Plotted Emotion count from data frame.

6.3 FEATURE EXTRACTION

A crucial step in analyzing and determining relationships between various things is feature extraction. Since the models cannot directly understand the audio data that is provided, we must transform it into a format that the models can understand. To do this, we employ feature extraction. The audio signal is a three-dimensional signal in which three axes represent time, amplitude and frequency. There are various ways to do feature extraction some of which are Chroma Deviation, Spectral Centroid,Entropy of Energy, Zero Crossing Rate, Mel Spectrogram and so on. In this project for feature extraction

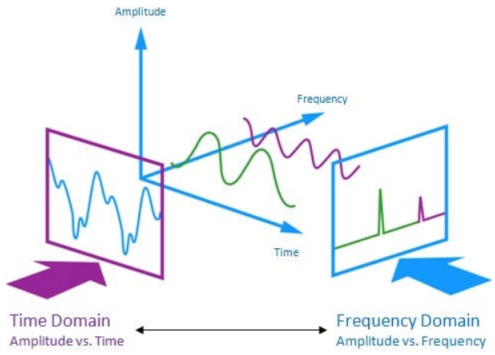

Figure 3: Three-dimensional audio signal.

Mel Spectrogram was used as we are using the CNN model to train our datasets. Spectrograms are generated with the help of Fourier Trans- forms on sound signals. The sound signal is divided into small segments of time to which FT is applied individually. Thus, we get a frequency versus time graph. Human perception of frequency, however, is logarithmic rather than linear. This issue is resolved by the Mel scale, which maps a tone's perceived frequency to its measured frequency.

6.4 TRANSFER LEARNING

The three databases are divided in an 8:1:1 ratio into training, validation, and testing sets. We use the training and validation datasets to optimize our Mel spectrogram data transfer learning models. For the use in ensemble stage, the confidence scores for the testing split dataset are created for each of the three models and saved as a CSV file. We used three transfer learning models Namely GoogleNet, Inception-v3, ResNet18

6.4.1 GOOGLENET. Google Net was proposed by research at Google (with the collaboration of various universities) in 2014 in the research paper titled "Going Deeper with Convolutions". The GoogLeNet architecture is very different from previous state-of-the-art architectures such as AlexNet and ZF-Net. It uses many different kinds of methods such as 1×1 convolution and global average pooling that enable it to create deeper architecture. In Google architecture,

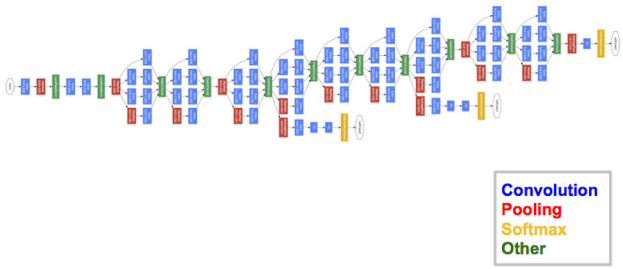

Figure 4: layer by Layer architectural details of GoogLeNet.

there is a method called global average pooling is used at the end of the network. This layer takes a feature map of 7×7 and averages it to 1×1. This also decreases the number of trainable parameters to 0 and improves the top-1 accuracy by 0.6The inception module is different from previous architectures such as AlexNet, and ZF-Net. In this architecture, there is a fixed convolution size for each layer. In the Inception, module 1×1, 3×3, 5×5 convolution and 3×3 max pooling are performed in a parallel way at the input and the output of these are stacked together to generate final output. The idea behind that convolution filters of different sizes will handle objects at multiple scales better.

6.4.2 INCEPTION-v3. Inception-v3 is a widely used 2D CNN model known for its strong classification performance through transfer learning. It's an extension of the GoogLeNet model, incorporating Batchnorm extensively in activation layers. Operating on input sizes of 299 × 299 × 3, it employs various convolution layers for feature extraction. The model's inception blocks enable diverse filter computation by concatenating them into a single feature map, reducing computational complexity by minimizing parameters.

6.4.3 ResNet18. ResNet, also known as Residual Network, was introduced in 2015 to tackle the issue of the "degradation problem" in deep networks by utilizing residual mapping. This architectural approach significantly improves the optimization process of CNN models. Similar to many other popular CNN models for image recognition, ResNet-18 is pre-trained on the ImageNet dataset, accepting input images of size 3 × 224 × 224, which is smaller than that of the Inception-v3 model. The performance of ResNet improves with increasing network depth. Variants of ResNet with different depths include ResNet-18, ResNet-34, ResNet-50, ResNet-101, and

ResNet-110. Among these, ResNet-18, utilized in our proposed three-stage-pipeline framework, strikes a balance between computational complexity and accuracy.

7 RESULTS AND DISCUSSION

Here we providing tabular data for results that we have obtained after working on the three different dataset. We give a thorough explanation of how we evaluated metrics, the CNN Transfer learning models' performance, and the final ensemble model. By comparing our work with previous research, we've shown that our method achieves the best performance for the SER problem

7.1 EVALUATION METRICS

To evaluate the performance of our proposed three-stage model. we have considered F1-Score, Precision, Recall, and Accuracy as our evaluation metrics. A majority of past researches have used Accuracy as the standard metric for the evaluation of performance. As a result, we will be providing a comparative study between our Three-Stage model and previous models for the SER problem. We can figure out the evaluation metrics by looking at simple things like True Positives, True Negatives, False Positives, and False Negatives. The corresponding formulas are as follows:

Accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{2}$$

Precision:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{3}$$

Recall:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{4}$$

F1 Score:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

We chose Precision, Recall, and F1 scores because our database has an uneven distribution of samples. In the next parts, we compare our SER model, which uses both deep learning and machine learning classifiers, with previous ones.

7.2 PERFORMANCE OF CONSTITUENT MODELS

Our all three constituent models has been loaded from the Pytorch Model Zoo and is pretrained on the ImageNet [2] dataset. The entire model weights have been freezed except the classification layers. The classification layers initially had an output layer with softmax activation of size (1, 1000) which are fine-tuned to (1, num-of-classes). All three has been trained for exactly 50 epochs on SAVEE, 25 epochs for CREMA-D, and 10 epochs for RAVDESS dataset after which the best validation accuracy has been taken into consideration. Adam optimizer has been used for gradient descent with the learning rate as 0.001 and values (0.9, 0.99). The training process has been experimented with different learning rates, batch sizes, and number of epochs and the final values have been experimentally chosen for the Three-stage model. The Inception-V3 model achieved 98.76 accuracy, ResNet18 model achieved 98.77% accuracy, and GoogLeNet model achieved 99.08% accuracy for the CREMA-D dataset. For the RAVDESS dataset, the Inception-V3, ResNet18, and GoogLeNet models achieved an individual classification accuracy of 92.07%, 95.29%, and 97.24% respectively. Figure 10 shows the learning curves for the SAVEE dataset using the Inception-v3, GoogLeNet, and Resnet18 models respectively. We can infer from the graphs that the model does not learn anything new and reaches a maximum accuracy around the 10th epoch. Similarly, RAVDESS dataset, the training of the model reaches a point where the accuracy halts to improve which is again around the 10th epoch. All the learning curves have been plotted using the TensorBoard library in Python.

8 FUTURE WORKS

In future work, we aspire to integrate the ensembling phase using the results that we got from the three pre-trained models in the transfer learning phase. Unfortunately, due to limitations in time and resources, we were unable to pursue this aspect in the current research. However, integrating the ensembling phase might play a great role in augmenting the performance and reliability of our model.

By combining the predictions from several base models into a single final prediction, ensembling allows the optimization of each model's advantages as well as minimizes its drawbacks. In future, Our goal will be to create a prediction mechanism that will be more reliable and accurate by ensembling the classification scores produced by the three CNN Transfer learning models that were previously discussed.

In our upcoming work, We plan to investigate the penalization of other class predictions and the assignment of the three-stage pipeline model classification scores in. We hope to improve this and produce more accurate prediction scores for the number of classes in each database. In addition, we intend to assess our model's efficacy using measures including F1 score, accuracy, recall, and precision. For every dataset, these metrics will provide a thorough understanding of how well our "ensemble model" performs across various emotion classes. Furthermore, we plan to investigate the use of binary classification techniques, including the One vs. All method, to improve the model's class discrimination capabilities.

We also want to research the receiver operating characteristic (ROC) curves for our ensemble model on the RAVDESS, CREMA-D, and SAVEE datasets in the near future. The ROC curves will provide valuable insights into the model's ability to differentiate between classes. We will aim to better understand the model's performance and pinpoint areas that still need work by analyzing these curves in-depth.

9 CONCLUSION

In this research, a new method for Speech Emotion Recognition (SER) using 2D CNN models and transfer learning is presented. This work highlights how important it is to use pre-trained models on large picture datasets in order to extract important characteristics from Mel spectrograms of audio data, thereby redefining the issue as a computer vision problem. Across benchmark datasets, our model, which uses a three-stage pipeline, achieves remarkable performance in reaching state-of-the-art accuracy. The three-stage

pipeline's dynamic ranking model makes it possible to make flexible predictions without having to reinitialize weights for every dataset, which increases the model's adaptability and effectiveness. Additionally, the design of the pipeline reduces the errors made by individual CNN classifiers, which enhances overall performance. The SAVEE, CREMA-D, and RAVDESS datasets show encouraging accuracy rates of 98.57%, 99.38%, and 99.66%, respectively, according to experimental results. Although, through this project we were able to get good accuracy there is still room for improvement by introducing ensembling. Other feature sets or data augmentation methods might be investigated in future studies so that we can improve the performance of the model.

REFERENCES

[1] Meaad Hussein Abdul-Hadi and Jumana Waleed. 2020. Human Speech and Facial Emotion Recognition Technique Using SVM. In *2020 International Conference on Computer Science and Software Engineering (CSASE)*. 191–196. https://doi.org/10.1109/CSASE48920.2020.9142065

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[3] Dias Issa, M. Fatih Demirci, and Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control* 59 (2020), 101894. https://doi.org/10.1016/j.bspc.2020.101894

[4] Eva Lieskovská, Maroš Jakubec, Roman Jarina, and Michal Chmulík. 2021. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. *Electronics* 10, 10 (2021). https://doi.org/10.3390/electronics10101163

[5] S. R. Livingstone and F. A. Russo. 2018. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English. *PLoS ONE* 13, 5 (May 2018), e0196391. https://doi.org/10.1371/journal.pone.0196391

[6] Cristina Luna Jiménez, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan Montero, and Fernando Fernández-Martínez. 2021. Multimodal Emotion Recognition on RAVDESS Dataset Using Transfer Learning. *Sensors* 21 (11 2021), 7665. https://doi.org/10.3390/s21227665

[7] Mustaqeem, Muhammad Sajjad, and Soonil Kwon. 2020. Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM. *IEEE Access* 8 (2020), 79861–79875. https://doi.org/10.1109/ACCESS.2020.2990405

[8] Peipei Shen, Zhou Changjun, and Xiong Chen. 2011. Automatic Speech Emotion Recognition using Support Vector Machine. In *Proceedings of 2011 International Conference on Electronic Mechanical Engineering and Information Technology*, Vol. 2. 621–625. https://doi.org/10.1109/EMEIT.2011.6023178

[9] Prabhav Singh, Ridam Srivastava, K. Rana, and Vineet Kumar. 2021. A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems* 229 (10 2021), 107316. https://doi.org/10.1016/j.knosys.2021.107316

[10] Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. Speech Emotion Recognition with Dual-Sequence LSTM Architecture. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6474–6478. https://doi.org/10.1109/ICASSP40776.2020.9054629