Multimodal Machine Learning for Emotion Recognition

by

Margulan Kazikhan

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

June 2025

Author .....
Department of Computer Science
June 2025

Certified by .....
Ben Tyler
Associate Professor, Department of Computer Science
Thesis Supervisor

Certified by .....
Adnan Yazici
Professor and Chair, Department of Computer Science
Thesis Supervisor

Accepted by .....
Yelyzaveta Arkhangelsky
Dean, School of Engineering and Digital Sciences

Multimodal Machine Learning for Emotion Recognition

by

Margulan Kazikhan

Submitted to the Department of Computer Science
on June 2025, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

Emotion recognition has become a popular research area in recent years due to the abundance of useful applications. This technology has been used in a variety of areas, including social media, crowd monitoring, live streaming, and human-robot interaction. Recent approaches to emotion recognition have used neural networks such as transformers, multimodal classification, LSTMs, and convolutional neural networks. Recent research has been facilitated by publicly available datasets, which include videos of persons that have been labeled with the dominant emotion of the given scene. In this work, a multimodal technique is used to classify scenes by emotional expressions from such videos by extracting video frames, audio, and transcribed text. In this work, we have investigated ways to achieve improved performance and efficiency at each stage of the classification process, where we have focused on developing and refining the preprocessing stages of each data input type. This work has allowed us to achieve 89% accuracy on a commonly-used dataset, using a combination of video, audio and text.

Keywords: Emotion Recognition, Multimodal Learning, Deep Learning, Image Processing, Intention Estimation.

Thesis Supervisor: Ben Tyler
Title: Associate Professor, Department of Computer Science

Thesis Supervisor: Adnan Yazici
Title: Professor and Chair, Department of Computer Science

Acknowledgments

I would like to express my gratitude to my supervisor, Prof. Ben Tyler, and my co-supervisor, Prof. Adnan Yazici, for their support, guidance, and encouragement throughout the course of this research. Their insights, feedback, and patience were invaluable to the development and completion of this thesis.

I am also grateful to the faculty and staff of the Computer Science department at Nazarbayev University, whose knowledge and assistance have contributed to my academic and personal growth.

Special thanks to my colleagues and friends, for their collaboration, motivation, and many helpful discussions during the research process.

Lastly, I would like to thank my family for their unwavering love and support. Their belief in me made this journey possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Emotion recognition has developed into a popular research area due to the large number of applications, which has been facilitated by the increased availability of data—particularly in the form of video content. Internet streaming, recordings, and video cameras have become primary sources of information that can be leveraged for a large variety of tasks, including emotion recognition. Videos and other data sources can be analyzed using machine learning techniques to estimate emotions that are being expressed by humans in the given scenario.

In this work, we present a multimodal technique to estimate emotions being expressed in video scenes by extracting video frames, audio, and transcribed text. Before going into detail, we will describe some areas where this can be applied in the real world.

1.1 Applications of the Work

The potential applications of this technology are vast and varied. For example, video cameras could be used in crowded public areas to provide large sets of human emotion data that can be interpreted by entrepreneurs, social organizations, and data analysis corporations. The emotional state of persons could potentially predict the choices and decisions made by the majority. This information can be leveraged for advertising, and other things such as social and political campaigns. A similar application could

be used in a sports arena, which could try to gauge spectators' moods and interests and determine the best camera shot for the viewers and provide the best angle for the advertisements. Analyzing the emotional state of the crowd in the stadium may help to automate this process [2].

Another application could be for private home owners who wish to technologically enhance their homes by using Internet of Things (IoT) devices. These devices could use emotional recognition systems to track the client's behavior and offer best solutions for their particular needs. Such devices can also be used in the home to track the state of the sick or elderly, and to alert the relevant persons if a medical intervention may be needed.

One of the most promising areas for this technology is in human-computer interaction, where understanding the emotional state of the user can greatly enhance the user experience. For example, virtual assistants, chatbots, and customer service robots could benefit from emotion recognition to tailor their responses based on the emotional tone of the user. In healthcare, emotion recognition systems could be used to monitor the emotional well-being of patients, providing valuable insights into their mental state. Additionally, the system could be applied in educational settings to gauge student engagement and emotional reactions, allowing for more personalized learning experiences.

1.2 Multimodal Approach to Emotion Recognition

Datasets in this domain are publicly available with the emotional labels provided. Data that can serve as input to emotion recognition systems include videos that are segmented into frames, audio recordings from the videos, and their text transcriptions. Multimodal approaches to emotion recognition use these distinct types of data, generally classifying them separately and concatenating the resulting vectors for further processing. Such approaches generally achieve higher accuracy results due to the embedding of external data in the preliminary stages [24]. Fusing the models in between with emotional embeddings allows for the parameters to be tuned along

with better interpretation of the already processed segments. Increased research in multimodal emotion recognition has focused on the abundance of methods that can be used for these additional embeddings.

In such approaches, the results are interpreted according to the needs of the classifier and extract the emotion labels based on the algorithm chosen. In general, the objectives of emotion recognition approaches can be divided into two categories: good performance, and high efficiency. The best performance models are oriented for the highest accuracy with little attention on how much time, resources, and computational power it is going to cost [10]. The high efficiency approaches are more time-oriented, and try to provide acceptable accuracy with little complexity, low computational power, while producing "good enough" results as early as possible.

The current view of ongoing research in the area can be divided into several parts. While transformer models have been successful in classifying emotions [25], more recent papers outperform those scores and propose models different from transformers [23]. Such models include convolutional neural networks [25]. Videos used as raw data generally have to be transferred into segments of frames, which align well with convolutional neural networks. The preprocessing part takes the frame and extracts human faces so that they can be fed into the network. Long Short-Term Memory models (LSTMs), intra-model and cross-model fusion, fully connected networks are additional types that are used the area of emotion recognition. More details are provided in the following chapter.

1.3 Novelty of the Approach

Multimodal learning shows promising results in the interpretation of human emotions [23]. Here, we are going to apply multimodal learning as the key feature in combining classification results for each data type. Videos are going to be segmented into sequences of frames, which are then classified using convolutional neural networks which are part of the large model structure. The audio recordings are going to be converted into spectrogram images and will be preprocessed before they are input

into the convolutional neural network. This network is another part of the large multimodal learning structure. The text transcription of the audio is going to be interpreted using a pre-trained large language model such as BERT [8].

The goal of this work is to improve the overall results of the emotion recognition task using the multimodal approach. The main novelty of this work is to apply custom preprocessing and postprocessing to improve upon the existing results. The strict order of the input from each data type allows us to achieve such higher accuracy scores.

The preprocessing approach applied to achieve better classification accuracy is a little different compared to other works. For example, most other works include a general preprocessing algorithm to transform the input data into a form to be used as input to the classification model. This preprocessing is different, focusing on each data type where the videos use custom algorithms that were created to serve the needs of the model, and the audio and text have their own processes that are used to achieve the best accuracy for them.

The postprocessing approach that we use is also a little different. Usually, the model output is fused directly into the stage of merging, where here, the model creates the numerical values that are assessed and evaluated using specific algorithm, and further decisions are made based upon that evaluation. All of this provides good results at the end of fusion.

1.4 Thesis Outline

The main points of the system can be summarized as follows:

- Use custom preprocessing to achieve higher efficiency
- Apply fusion architecture to classify different data types simultaneously
- Combine models with multimodal learning technique to optimize performance
- Test model performance using different postprocessing approaches

In Chapter 2 we will discuss the existing literature and how it relates to this work. In Chapter 3 the methods used in the work are presented in more detail, along with the overall system architecture. Chapter 4 covers the primary experiments that were run and the results of the research work. Chapter 5 concludes the work and includes a brief discussion of the possible ways to improve the model.

Chapter 2

Related works

The task of emotion recognition is demanding in terms of computational resources. Simple analysis is usually not enough to accurately determine the expressed emotion, and generally requires a more thorough analysis. Early attempts in emotion recognition include analysis of different types of data, and researchers were analyzing faces, audio, and text independently from each other. The models included Support Vector Machines, k-Nearest Neighbors, and Gaussian Mixture Models. For text keyword spotting, lexicons and basic sentiment analysis have been used [7]. Generalization techniques included average scores and rule-based decisions [7].

The latest state-of-the-art approaches for emotion recognition utilize multimodal solutions. Multimodality is applied to combine separate classifications of different data types such as audio, video, and text. For image classification, convolutional neural networks are used. Generally, an adapted model from previous research work is used, where works that show the highest performance use a custom design of the convolutional network, such as [18] described below. The audio files are transformed into audio waveforms that are interpreted as images. In most of the cases the transformation is performed using Short-Time Fourier transform and Mel-Frequency Cepstral Coefficients [6]. The text is classified using large language models such as BERT [8]. All of these are combined using various multimodal techniques.

The state-of-the art results were achieved in the work by H-D. Le, G-S. Lee, S-

H. Kim, S. Kim, and H-J. Yang [18]. In this work the accuracies of 85.9% on the IEMOCAP dataset [4] and 67.8% on the CMU-MOSEI dataset [27] were shown. The large model structure consists of several models combined into the multimodal architecture. Convolutional neural networks were used for image classification, while the pre-trained BERT model was used for text classification. The classification models were combined into the transformer, which produced preliminary results that were processed later into fully connected layers. Emotional embeddings are then applied in the last stage.

Another approach considers adding small and lightweight models into the classification system. In the study done by A. Radoi and G. Cioroiu [23], a system that relies on energy efficiency is employed. The convolutional neural networks are used both for image classification and audio classification tasks. Audio is converted into the waveform spectrograms beforehand using the Librosa library [20]. The step-by-step incremental approach allowed to reach accuracies of 76.3% on the RAVDESS dataset [19] and 74.2% on the CREMA-D library [5]. This solution can be suitable for situations such as online emotion recognition.

The work presented by D. Valles and R. Matin [26] trained their models on a combination of datasets, which were classified further using the ensemble learning technique. This type of classification uses a unique interpretation of outputs that were produced by distinct models. The accuracy of 66.5% on the RAVDESS, TESS [11], and CREMA-D datasets with added background noise was reached.

A different approach often used by researchers in healthcare is to use medical data measurements taken from the human body, which are then used as direct input. The electrocardiograms (ECGs) are often chosen as a source of model data. The work done by M. A. Hasnui, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz [13] achieved the accuracy of more than 90% using this approach. The ECG is applied in combination with other data types which resulted in high accuracy scores which are competitive with state-of-the-art solutions [13].

The work done by S. Akbar, A. Raza, T. Shloul, A. Ahmad, A. Saeed, Y. Y.

Ghadi, O. Mamyrbayev and E. Tag-Eldin uses ensemble learning to combine models [1]. The difference of ensemble learning from multimodal architecture is in flexibility during training. The training process in a multimodal architecture is sequential, where the data at preliminary stages can be analyzed to adjust parameters. The ensemble learning architecture provides for continued training once the input is received and the process finishes with the final output. The ensemble learning approach has a larger variety of ways in which models can be combined. The mentioned research achieves the accuracy of over 90%.

Other common solutions use models like Support Vector Machines, Long-Short Term memory, and a combination of models. In the work done by S. Harikant, V. Lakshmi R. and R. Prasad [12] the Support Vector Machine showed an accuracy of 74.3% on the RVM [12] emotion database. In the Korean research done by F. M. Talaat [14], the convolutional neural network and long short-term memory models combined showed an accuracy of 96% on a domestic dataset collected using university resources. Another ensemble model by P. Dkhara, P. K. Singh and M. Mahmoud [9] performed 90.84% on the DEAP [16] and AMIGOS datasets [21]. The complex preprocessing stage allowed the convolutional neural network to reach an accuracy of 99.99% [25]. The preprocessing stage converted audio files into spectrogram images using a specific approach done by Jo. A. Hypson [25].

A summary of these related works is shown in Table 2.1.

This work uses the IEMOCAP dataset that was also used in the state-of-the-art work mentioned above [18]. In that research study, the combination of convolutional neural networks, BERT, transformer, and fully connected layer resulted in the accuracy score of 85.9%. This work uses a similarly structured classification model, and reaches an accuracy score of 89.1%.

This work uses a multimodal approach, where the models used for video, audio, and text are close to the state-of-the-art solutions, but our preprocessing stage is different from those works presented in this chapter. The data types are processed independently using a custom algorithm. The scene analysis allows the system to

focus only on the necessary part of the data. For example, the frame cut is applied to the scene whereas other papers use face cut or avoid cutting altogether. The processing of inner stages in the multimodal structure was applied using an approach that is different from the works presented in this chapter as well. The manual nature of the algorithm allowed us to dive deeper into the data.

Paper	Model	Dataset	Accuracy
"An Audio Processing ... with ASD" [26] , D. Valles, R.Matin	MLP, SVM, RNN, Ensemble Model	RAVDESS + TESS + CREMA-D + Noise	63-66%
"Uncertainty-based ... Recognition" [23] , A. Radoi, G. Cioroiu,	Multimodel	RAVDESS, CREMA-D	74-76%
"Multi-Label ... Learning" [18] , H-D. Le, G-S. Lee, S-H. Kim, S. Kim, H-J. Yang	CNN, BERT, Transformer, FC	IEMOCAP, CMU-MOSEI	68-86%
"Speech ... Learning" [3] , H. Aouani, Y. B. Ayed	SVM, Encoders	RVM	74%
"Real-time ... IOT", F. M. Talaat [25]	CNN	IOT	100%
"A fuzzy ... Recognition" [9] , P. Dkhara, P. K. Singh, M. Mahmoud	CNN, LSTM, GRU, Ensemble Model	DEAP, AMIGOS	91%
"Speech Emotion ... Information" [14] , A-H. Jo, K-C. Kwak	Bi-LSTM, CNN, Two-Stream Model	KSERD	90-96%

Table 2.1: Related Literature

Chapter 3

Methodology

This work consists of several stages: the data phase, preprocessing, classification, postprocessing, and fusion. Each phase has its own processes divided between the different data types. In some of these, the merging of some data types is performed, while in others the data types are considered separately. An overall picture of the system's architecture can be seen in Fig. 3-1.

3.1 Data Phase - The IEMOCAP Dataset

The initial part of the whole process starts in the data phase. The IEMOCAP dataset is used in this study, which contains data from 10 speakers, both male and female [4]. IEMOCAP consists of 151 videos of pairs of actors that express emotions looking in the forward direction. The dataset includes approximately 8,000 training samples and 800 test samples, spanning five sessions. Emotions are labeled using sensors and manual annotations, with categories including disgust, excited, other, xxx, neutral, happy, sad, frustrated, surprised, excited, fear and angry. The xxx label indicates that the emotion was failed to be labeled by sensors and human judges.

The diversity of the IEMOCAP dataset ensures that the model is exposed to a wide range of speech patterns, accents, and emotional expressions. This diversity is critical for developing a robust emotion recognition system that generalizes well

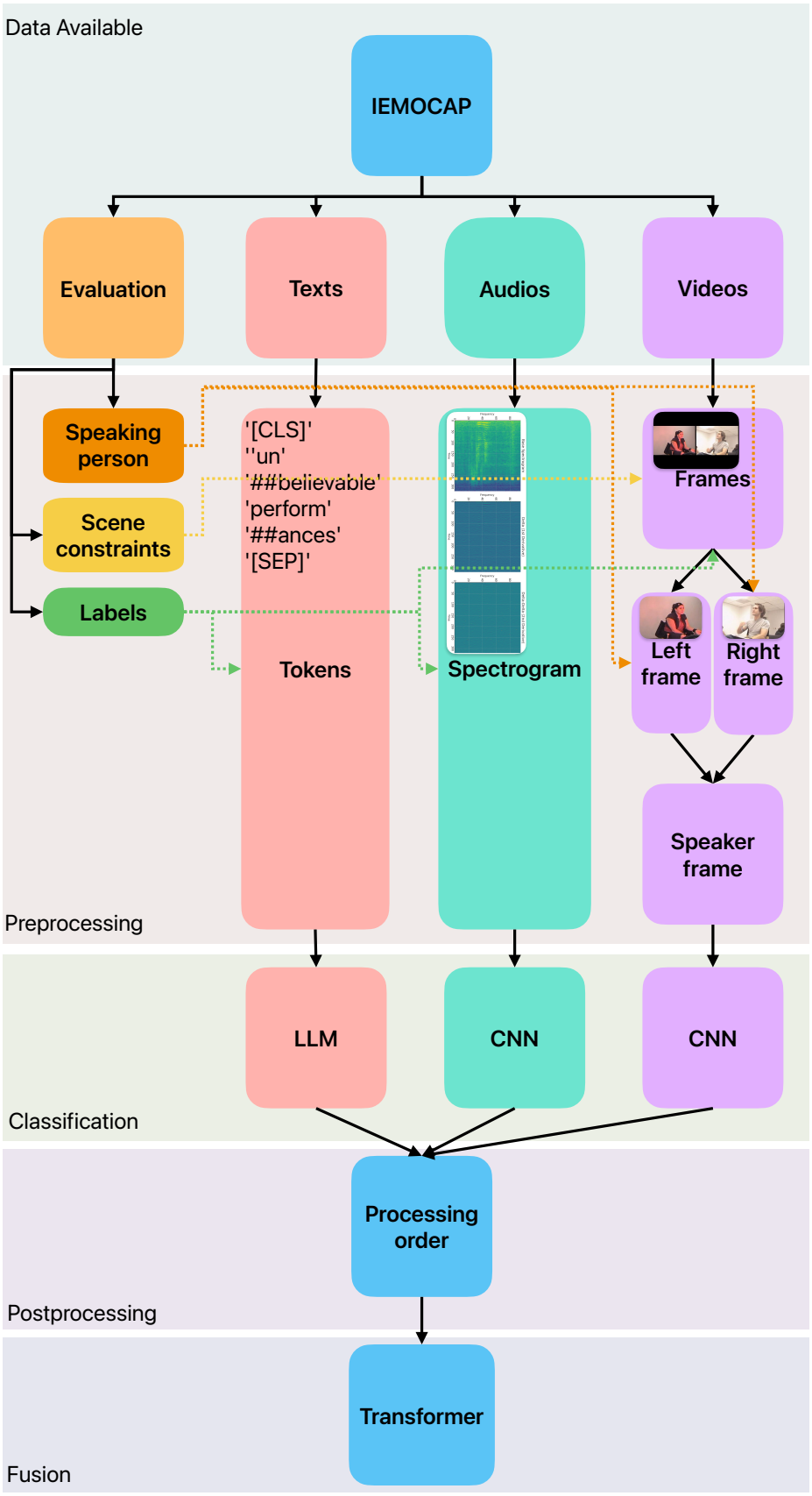

Figure 3-1: Architecture

across different user demographics. Furthermore, the dataset includes multimodal annotations, linking audio and text data for each utterance. This alignment facilitates the integration of text and audio features during the model training process, ensuring consistency and coherence in emotion classification.

The data IEMOCAP provides is fairly comprehensive and we make use of much of its features in this work. The videos are the primary source of information, with the audio being part of the videos, though the audio recording is saved separately for use. The audio part was analyzed for words and saved separately for use as transcriptions. An important part of the dataset is the data from sensors, human labeling and scene position information. Here, we use the scene code to determine which person in the scene is speaking, the scene constraints to help in extracting frames from video, and emotion labels for training and evaluation of the system models.

3.2 Preprocessing

The next stage is the preprocessing phase. Preprocessing is divided between the video, audio, and text data types, as well as an evaluation part which is adjacent to all other parts. The label extracted from the evaluation part is provided to all data types at the end of the preprocessing phase.

The video preprocessing consists of video frame extraction and cropping. Video frame extraction is the process of dividing video into slices of images. Each slice is captured within the specific time frame that is specified beforehand by the scene constraints. The time frame looks like [3.05 - 7.06]. The video is made of frames, with the frame rate of each video being 30 fps. Based on that information the video timeline is divided into the frame sequence, and the time frame is divided into the frame interval. The frame interval looks like [372 - 467]. The middle frame is calculated as the average of those numbers and extracted. The video is essentially sliced into time interval frames providing the source of images that are going to be processed later.

The frame is then divided into left and right parts, as the scenes in the IEMOCAP dataset consist of two speakers on the left and right. The scene actor always sits on the left side, and the supporting actor always sits on the right and is of the opposite gender of the scene actor. The scene code is used to determine the gender of the scene actor. For example, suppose we take the scene code of "Ses01M_impro01_F003". The letter at index 5 can be "M" and "F", where "M" indicates that the scene actor is male, and "F" indicates that the scene actor is female. Here, the character is "M", which means that the scene actor is male and is sitting on the left, and the supporting actor is female and to the right. The last subword in the scene code indicates the speaking person, who is the person whose emotion is labeled. Here, the starting letter "F" of the last subword indicates that the speaker is female, who is the supporting actor in this case. In any case, the frame is cropped so that the side of the frame containing the speaker is used in the subsequent stage.

The preprocessing part is also connected to the audio transformation. There are a couple of steps in the conversion to a spectrogram image. First, the wave and frequency information is obtained. Then, the audio files are processed to produce the spectrogram images containing the sound frequency and intensity. Each audio file is connected to the corresponding video frame. The goal of the classifier is to tell whether the particular video frame has a certain emotion in it later. For that reason, the audio interval that has the corresponding video frame connected to it is converted separately to give more information later. The waveform spectrogram image is obtained using the Librosa library [20]. This library allows us to obtain a spectrogram that is different from short-time Fourier transform and Mel-Frequency Cepstral Coefficients.

The text preprocessing is done using the algorithm proposed by the BERT large language model [8]. The use of the language model here is to transform the text into a set of tokens, which is derived directly from the BERT model functions. Some of the tokens represent words and their parts. Other tokens represent flag rules for the

model; for example "[SEP]" means separate, where the given portion of tokens needs to be separated. The emotion labels to those tokens are assigned by the evaluation part based on its corresponding scene code.

3.3 Classification

The classification system consists of several parts as well. The first stage is classification of the video frames using convolutional neural networks. In this work, the video frame images are fed into the convolutional neural network. The design of the network is custom, based on the ResNet-18 model with trained weights. The balance between the lightweight and large system is found, using the fully connected layer to adjust the model to current needs. The Adam optimizer is the best option for this task in this situation [15]. The output for this model is the feature vector containing values corresponding to each of the labeled emotions. This vector is used further in the next phase.

The audio classification process uses the waveform spectrogram images from the previous phase, feeding them into another convolutional neural network. Note that the spectrogram waveforms are images of the same size. The design of the network is chosen to be based on AlexNet [17]. There are additional layers that would make the model suitable for input and the desired output. The custom design helps to find a balance between the lightweight and complex model for the best performance. The audio waveforms are matched with the corresponding video frame images to provide the feature vector as the output to be used as input for the next classification phase.

The text classification component uses a pretrained neural network, as large language models are difficult to customize for specific tasks. Here, we use BERT as our model, as that is most effective in transcription interpretation [8]. The transcription belongs to the video, and is divided into the parts which match the corresponding video frame and the corresponding audio file. The output of this model is again the

feature vector containing values corresponding to each of the emotions. This is going to be processed and combined with the other vectors from this phase as input in the next phase.

3.4 Postprocessing

The processing order is the key distinguishing postprocessing operation of this work. The feature vectors, labels, and scene codes are derived from the previous stages. The scene code is used in determining the ordering. The scene code of the frame, spectrogram, and text is taken, and is sorted as given in the dataset along with the corresponding feature vector and label. In those data types that miss the particular scene code, the vectors are skipped until all data types match together. Then the next scene code is analyzed for siblings. The goal of the algorithm is to connect the scene to its visual representation, sound representation, and transcription. The next model should see all sides of the same scene from all angles simultaneously to provide the best output. A sibling is the vector and label of the data type that matches the vector and label of other data types by the scene code. All scenes find their visual, sound, and text representations at this stage. The output of this stage is an ordered sequence of scenes with feature vectors of all data types and labels, which are then fed into the fusion model in the final stage.

3.5 Fusion

Transformer fusion is the technique that allows us to combine the feature vector outputs of the data type models from the previous classification stage. This approach was chosen, as the performance of this technique for this task is generally high [18]. The structure of the transformer fusion model consists of fully connected layers, transformer layers, and a concatenation of feature vectors that are positioned in parallel. The triple channel input takes a matrix of 3 data types by 11 emotion labels. The model provides an output as a certain emotion type.

Chapter 4

Results

In this chapter, we test the system described in the previous chapter on multiple configurations to find which gives the best results. In our trials, we use the IEMOCAP dataset with its video, audio, and text transcriptions of the recordings. Ready access to IEMOCAP for the work was officially given by the University of Southern California, the host of the dataset.

4.1 Video Classification

4.1.1 Classification using Faces Only

This section presents the performance of different models trained for emotion classification using only the faces from the IEMOCAP dataset. Various architectures, including small CNN models and larger deep learning models, were evaluated on validation accuracy. Results are given below, where the overall performance indicates the model results for both actor faces in the scenes. The left face accuracy indicates the performance on only the left screen actor's face, and the right face accuracy indicates the performance on the right screen actor.

We see that the small convolutional neural network model exhibited relatively low performance in all cases:

- Overall validation accuracy: 29.8%
- Left face validation accuracy: 27.8%
- Right face validation accuracy: 29.6%
- Balanced dataset validation accuracy: 29.2%

These results indicate that the small model struggled to learn discriminative features effectively.

Performance improved significantly when using larger models, where the best results were obtained with ResNet-18:

- ResNet-18 overall validation accuracy: 65%
- ResNet-18 on left face: 96%
- ResNet-18 on right face: 92.8%
- ResNet-18 on balanced dataset: 94.3%
- Resnet-50 overall validation accuracy: 25%
- Resnet-50 on left face: 23.4%
- Resnet-50 on right face: 23.8%

Here, we see that the larger model (ResNet-18) significantly outperformed the smaller CNN model, suggesting the importance of deeper architectures for emotion classification. Performance varied based on whether the left or right side of the face was used for validation. ResNet-18 demonstrated particularly high accuracy when trained on the left face (96%) and right face (92.8%), whereas ResNet-50 failed to generalize effectively. Furthermore, the results for ResNet-50 on individual facial sides (23.4% and 23.8%) suggest potential overfitting or suboptimal feature extraction.

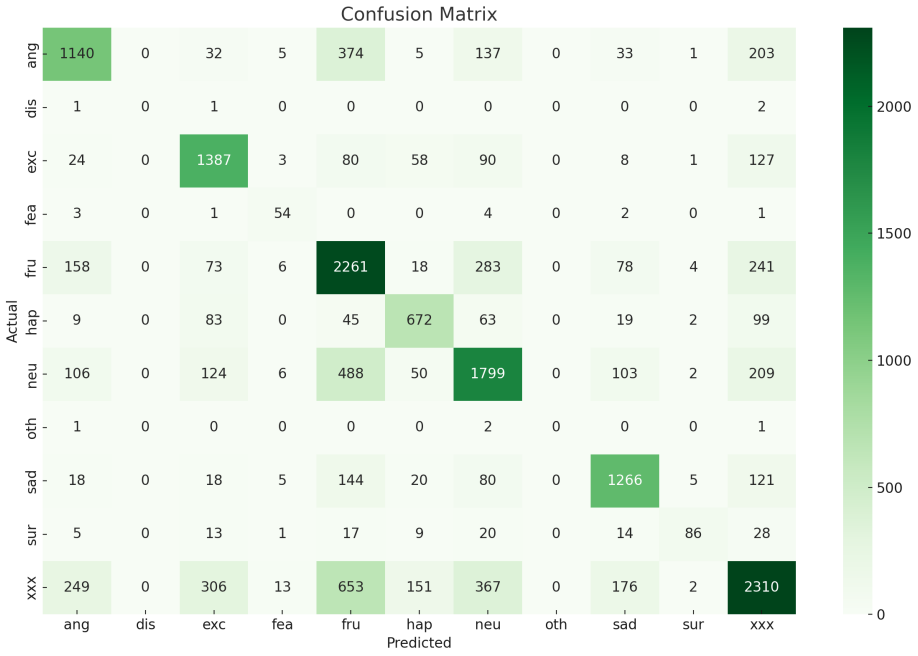

Figure 4-1: ResNet-18 overall confusion matrix

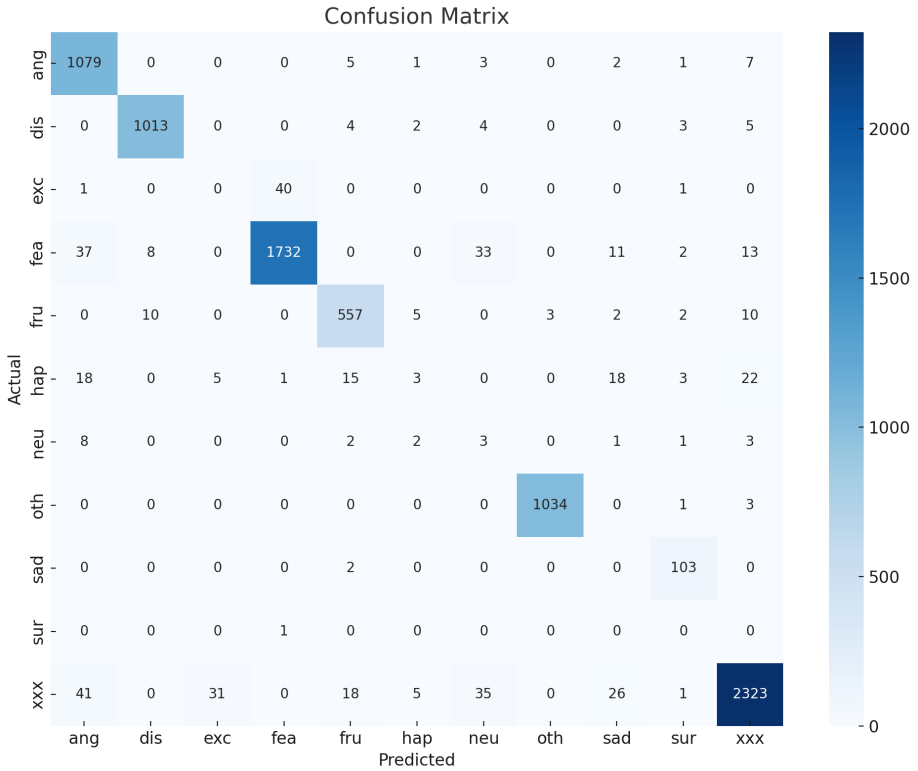

Figure 4-2: ResNet-18 on left face confusion matrix

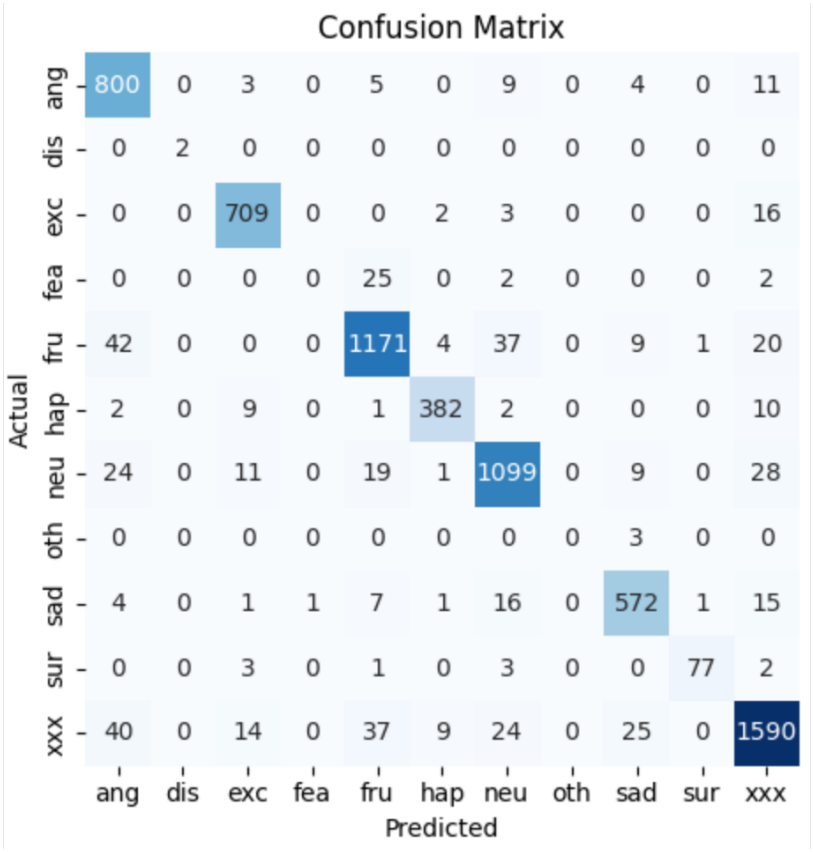

Figure 4-3: ResNet-18 on right face confusion matrix

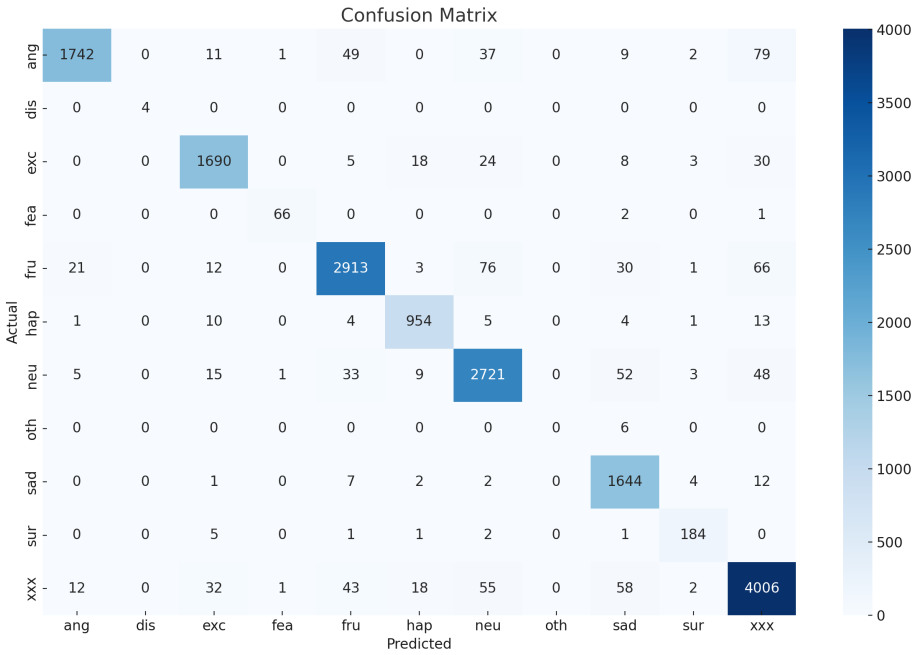

Figure 4-4: ResNet-18 on balanced dataset confusion matrix

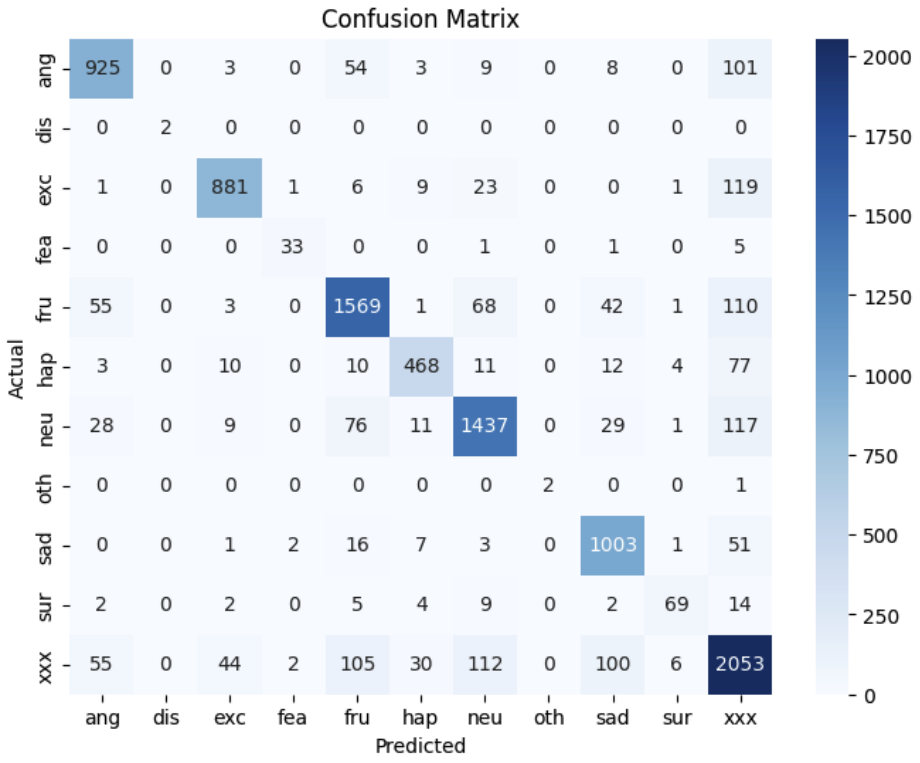

Figure 4-5: The ResNet-18 on speaker frame confusion matrix

Future improvements can explore fine-tuning the model hyperparameters and incorporating additional regularization techniques to further enhance generalization.

4.1.2 Classification using Frames

The frame classification that is chosen over face classification uses similar models. The logic behind this idea is that the body pose and hand movement provided in the frame give additional information for that. The frames are extracted and divided into right and left, and the speaker frame is chosen and goes to the ResNet-18 model.

ResNet-18 is trained on frames with similar characteristics. Here, the resulting accuracy exceeded 80%. The training of frames took more time than training of faces, given the increased number of features provided in the frames. In this case, only the speaker frame is classified. The training graph for that can be seen in Fig. 4-6.

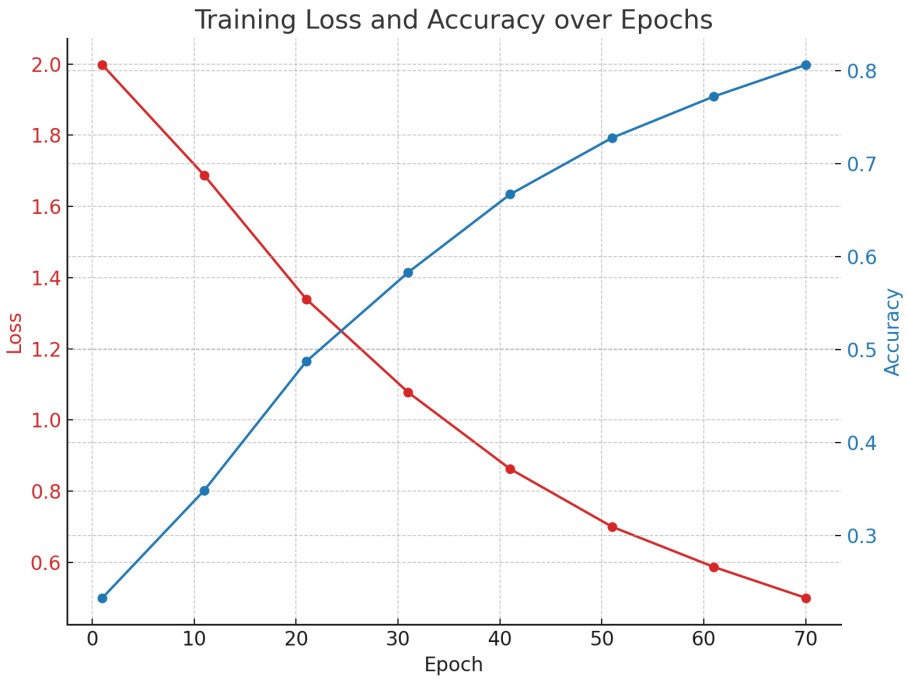

Figure 4-6: The graph for ResNet-18 on speaker frame

4.2 Audio and Text Classification

The proposed methodology combines text and audio processing pipelines, each utilizing a distinct model for feature extraction. The extracted features are fused to form a unified representation, which is then classified into emotion categories using a transformer. The feature extraction techniques for each modality are independently applied, followed by their fusion into a unified representation. This consolidated feature vector is then classified into one of several emotion categories using a transformer. The approach capitalizes on the synergy between text and audio data to improve the accuracy and generalization of the emotion recognition system.

The methodology involves several key steps. First, the sentences in the IEMOCAP dataset are tokenized using the BERT tokenizer, transforming the text into a format suitable for processing by a deep learning model. A pre-trained BERT model is then used to generate contextualized embeddings for each sentence. These text embeddings are further fine-tuned on the IEMOCAP dataset to improve their performance

for emotion classification. In parallel, the audio waveforms are processed by Librosa to generate Mel-frequency Cepstral Coefficients (MFCC), which are a common feature extraction technique for speech data. These spectrograms are then fed into AlexNet, a convolutional neural network (CNN) designed to extract relevant features from the spectrogram images.

Once both text and audio features have been extracted, the generated embeddings are concatenated into a single vector representation. This combined feature vector is then passed through a transformer and fully connected layer, which performs the final classification task. The transformer and fully connected layer maps the concatenated features to the predefined emotion categories (e.g., neutral, happy, sad, angry, etc.) The output of this layer is the predicted emotion label for the given input, providing an estimate of the speaker's emotional state.

For the implementation of the deep learning components, PyTorch is used as the primary framework. PyTorch is a powerful and flexible library for deep learning that provides efficient tools for training and fine-tuning models [22].

The fusion technique takes the feature vector output of those models and passes it into the next stage. The next stage is the transformer model that does the following operations of classifying feature vectors and comparing the output to the ground truth label.

The model demonstrated impressive performance, achieving an accuracy of 82% on the test set. Performance remained consistent across multiple runs. Notably, the training accuracy exceeded the test set accuracy, a common occurrence due to potential overfitting to the training data. However, the test set results remain strong, showcasing the model's ability to generalize well to unseen data. The confusion matrix can be seen in Fig. 4-7. Audio data alone showed lower accuracy compared to text data, emphasizing the importance of multimodal fusion.

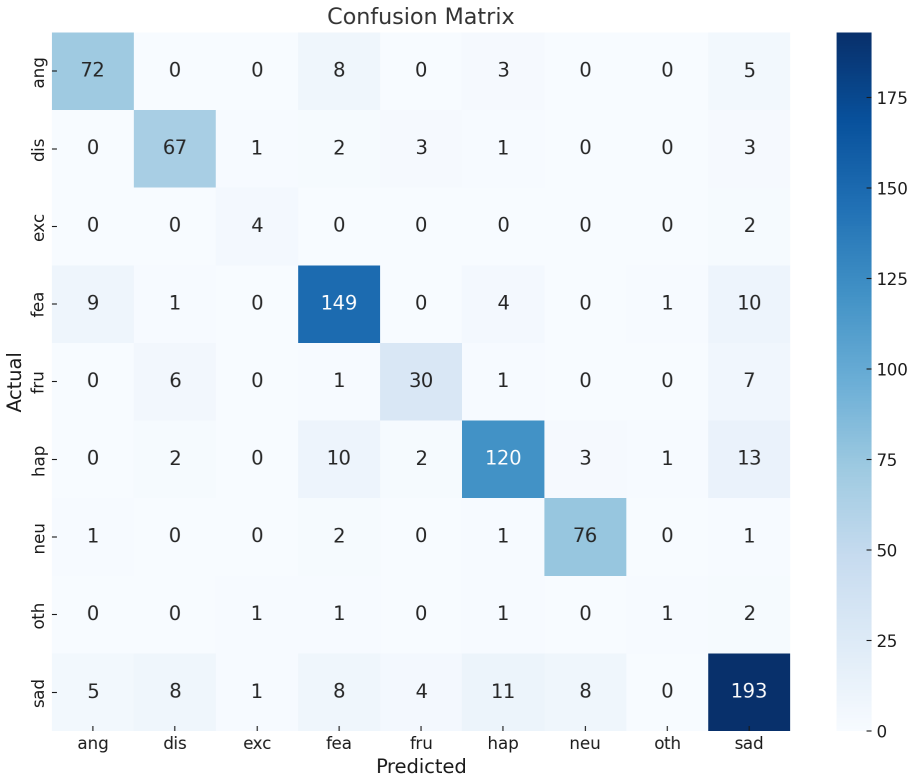

Figure 4-7: The resulting confusion matrix of audio and text

These results place the model at a competitive level compared to state-of-the-art methods, which typically report around 85% accuracy in similar emotion recognition tasks. Despite slightly lower accuracy, the proposed methodology's performance is still highly effective, indicating that the model can reliably classify emotions from both text and audio data.

4.2.1 Analysis of the Separate Models

A key strength of the model is its consistency across multiple training runs, with the model achieving peak accuracies on several occasions during the training process. This stability suggests that the model is well-optimized and not subject to significant fluctuations or instability in training, a crucial factor when developing models for practical applications. The training process was robust, and the model demonstrated solid convergence during the optimization phase, which contributed to its reliable

performance in emotion recognition tasks.

When analyzed separately, the audio and text data presented different levels of accuracy. The audio data alone demonstrated lower accuracy compared to the text data, emphasizing the importance of incorporating both modalities for optimal performance. While speech conveys significant emotional content through tone, pitch, and rhythm, the text provides critical semantic and linguistic information that helps contextualize the emotional expression. The lower performance of the audio modality in isolation highlights that audio data, while important, does not fully capture the emotional state without the complementary linguistic context that the text modality provides.

One possible direction for improving this approach is further fine-tuning the pre-trained BERT and AlexNet models to optimize their performance specifically for emotion recognition tasks. Additionally, exploring more sophisticated fusion techniques—such as attention mechanisms or temporal modeling to capture the dynamics of emotional speech could further enhance the system's robustness and accuracy.

4.3 Fusion: Frames, Audio and Text

The fusion that takes the post-processed data from previous stages is based on a transformer. Feature vectors are concatenated in parallel in the input, and then passed to the transformer, which finally gives the emotion derived from the system. The results of the transformer model, which represents the results of the whole multimodal architecture, are 89% in terms of accuracy. These results are better than those given in [18] by more than 3%, making it a candidate to become the new state-of-the-art for the IEMOCAP dataset.

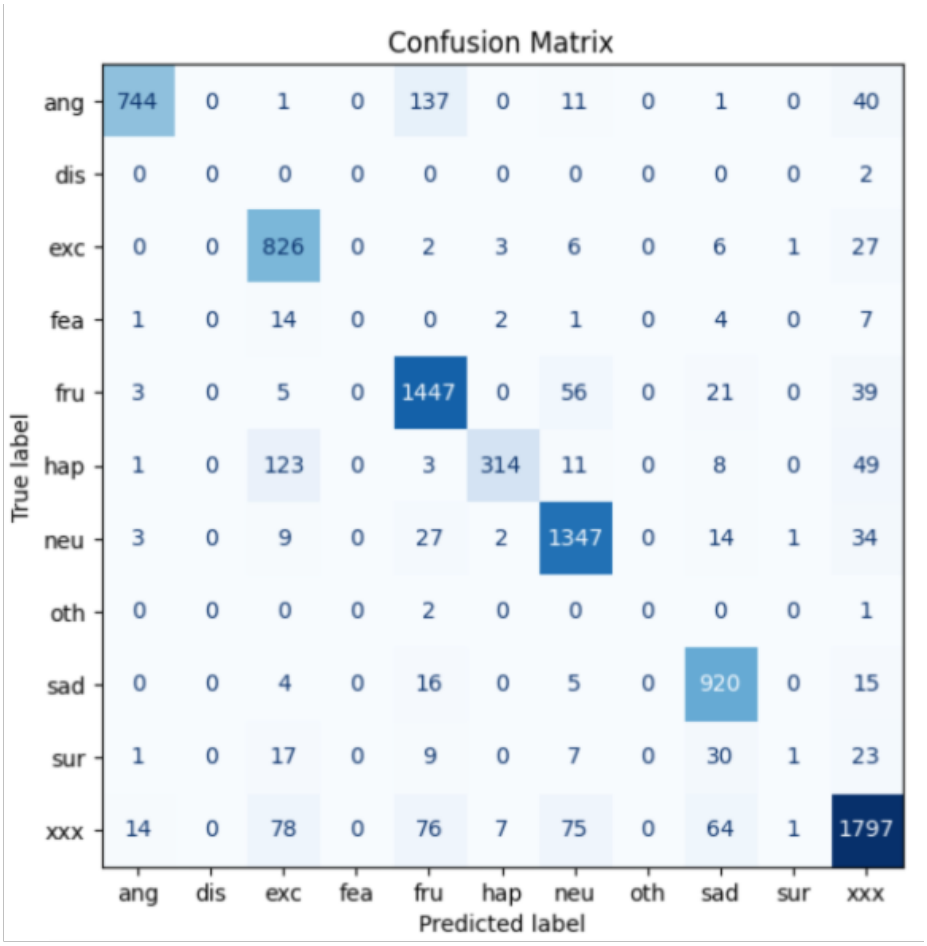

Figure 4-8: The transformer confusion matrix

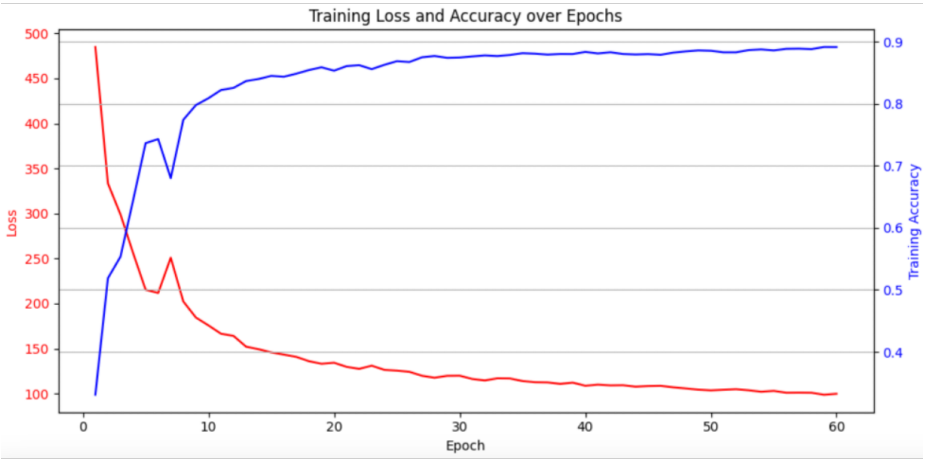

Figure 4-9: The graph for transformer

The primary influence on the results were the preprocessing and postprocessing algorithms. The preprocessing was customized to each data type. The video preprocessing was done using a unique approach, where the frame is extracted using the algorithm, and then cropped and passed through selection. The audio spectrogram was obtained with triple channels. The text preprocessing followed the BERT algorithm directly. The other novelty to the approach is in the postprocessing phase. Here, feature vectors obtained from the models are ordered exactly to match their scene. The final fusion model analyzes the scene while interpreting all aspects simultaneously.

Chapter 5

Conclusion

This work proposes a new model architecture in the field of emotion recognition. Its purpose is to make improvements in accuracy for the emotion recognition task. There are multiple strategies that are applied to accomplish this. Custom preprocessing by choosing the frame side is part of the strategy in optimizing the data that is processed in the neural network. Other parameters are tuned at the postprocessing stage. The model combination uses a custom neural network structure, and the transformer fusion technique is the focal point. The separate data type models showed peak accuracies of 96% for the face and 82% for audio and text. The resulting accuracy score of multimodal fusion is 89%. The proposed solution tends to work well with limited resources, which is good for users who do not have high-performance computing resources. This balance between performance and optimization allows the solution to be implemented and used in different areas of academia and industry.

Regarding the choice of video frames for processing, the option of time period estimation in the scene is an area for further research. Also, audio files can be transformed into spectrogram images in a variety of ways, which could be further investigated for the task at hand.

Bibliography

[1] S. Akbar, A. Raza, T. A. Shloul, A. Ahmad, A. Saeed, Y. Y. Ghadi, O. Mamyrbayev, and E. Tag-Eldin. patbp-enc: Identifying anti-tubercular peptides using multi-feature representation and genetic algorithm-based deep ensemble model. *IEEE Access*, 11:137099–137114, 2023.

[2] O. O. Et al. Multimodal emotion recognition using deep learning: A comprehensive study. *BMC Psychology*, 12(95), 2024.

[3] H. Aouani and Y. B. Ayed. Speech emotion recognition with deep learning. *Procedia Computer Science*, 176:251–260, 2020.

[4] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[5] Houwei Cao, David G. Cooper, Michelle K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

[6] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[7] C. M. de Melo, P. Carnevale, and J. Gratch. The effect of expression of anger and happiness in computer agents on negotiations with humans. *Pattern Recognition*, 45(10):4513–4521, 2012.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019.

[9] Trishita Dhara, Pawan Kumar Singh, and M. Mahmud. A fuzzy ensemble-based deep learning model for eeg-based emotion recognition. *Cognitive Computation*, 2023.

[10] C. Dixit and S. M. Satapathy. A customizable framework for multimodal emotion recognition using ensemble of deep neural network models. *Multimedia Systems*, 29(6):3151–3168, 2023.

[11] Shannon-Kay Dupuis and Kristine Pichora-Fuller. Toronto emotional speech set (tess). https://tspace.library.utoronto.ca/handle/1807/24487, 2010. University of Toronto, Department of Psychology.

[12] S. Harikant, R. Prasad, R. V. Lakshmi, and S. H. Speech emotion recognition using deep learning. *International Research Journal of Computer Science*, 9(8):267–271, 2022.

[13] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz. Electrocardiogram-based emotion recognition systems and their applications in healthcare—a review. *Sensors*, 21(15):5015, 2021.

[14] A-Hyeon Jo and K.-C. Kwak. Speech emotion recognition based on two-stream deep learning model using korean audio information. *Applied Sciences*, 13(4):2167, 2023.

[15] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2015.

[16] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, Jan.–Mar. 2012.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 1097–1105. Curran Associates, Inc., 2012.

[18] H.-D. Le, G.-S. Lee, S.-H. Kim, S. Kim, and H.-J. Yang. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning. *IEEE Access*, 11:14742–14751, 2023.

[19] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLOS ONE*, 13(5):e0196391, 2018.

[20] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. *Proceedings of the 14th Python in Science Conference (SciPy)*, pages 18–25, 2015.

[21] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. AMIGOS: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2021.

[22] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8024–8035, 2019.

[23] A. Radoi and G. Cioroiu. Uncertainty-based learning of a lightweight model for multimodal emotion recognition. *IEEE Access*, 1:1–1, 2024.

[24] T. N. Rincy and R. Gupta. Ensemble learning techniques and its efficiency in machine learning: A survey. In *2nd International Conference on Data, Engineering and Applications (IDEA)*, pages 1–6, 2020.

[25] F. M. Talaat. Real-time facial emotion recognition system among children with autism based on deep learning and iot. *Neural Computing and Applications*, 35:12717–12728, 2023.

[26] D. Valles and R. Matin. An audio processing approach using ensemble learning for speech-emotion recognition for children with asd. In *2021 IEEE World AI IoT Congress (AIIoT)*, pages 55–61, 2021.

[27] Amir Zadeh, Paul Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2236–2246, Melbourne, Australia, 2018. Association for Computational Linguistics.