Multimodal Emotion Recognition

Andreas Kouridakis
Postgraduate studentt
University of Piraeus
NCSR Demokritos
andreaskouridakis@outlook.com

Aris Tsilifonis
Postgraduate studentt
University of Piraeus
NCSR Demokritos
arisfo1998@gmail.com

Abstract — Emotion recognition from multimodal sources refers to the process of identifying human emotions using multiple types of data, such as speech, text, and facial expressions. This study investigates the use of machine learning models for emotion recognition, focusing on a six-class classification problem to identify six distinct emotional states. While other research often uses Mel-Frequency Cepstral Coefficients (MFCC) for feature extraction, we decided to create hand-crafted features for both speech and text data. Our study has two primary objectives: first, to evaluate the performance of these machine learning models when using only text data, only speech data, and a combination of both; second, to compare the performance of these machine learning models with some deep learning models [8],[9],[10] that have been used for emotion recognition using the same dataset as ours. Our findings indicate that combining text and speech data significantly enhances model's performance, and that machine learning models achieve accuracy close to that of deep learning models while being less complicated and more computationally efficient.

Keywords — multimodal emotion recognition, speech , text, machine learning

I. INTRODUCTION

According to the European Data Protection Supervisor, emotion recognition is the technology that analyzes signals such as facial expressions, voice tones, and body language in order to reveal information on one's emotional state. By combining data from these diverse sources, we can create a more accurate and reliable approach for detecting emotions. However, the challenge of accurately recognizing emotions stems from the complexity and variability of human emotional expressions, which can be influenced by cultural, contextual, and individual factors. Additionally, integrating data from multiple sources is complex, ensuring synchronization between different signals can be difficult. Previous research have demonstrated the benefit of using multimodal data in emotion recognition tasks and has identified various techniques for generating robust multimodal features. [1],[2],[3]. Some applications of emotion recognition are found in health care [4], [5], customer service [6] and education[7].

This paper is structured as follows: We begin with an overview of existing research in the field of emotion recognition in the Related Work section. The Dataset section describes the dataset used in our study, as well as the preprocessing steps undertaken. We also discuss the feature extraction process for both speech and text data. In the Machine Learning Models section, we outline the models employed in our experiments. The Metrics section defines the metrics used to evaluate our models' performance, explaining their relevance and providing insights into model accuracy and efficiency. The Experiments and Results section presents the experimental setup and results obtained, including an analysis of performance improvements achieved through the combination of text and speech data. Finally, the Future Work section discusses potential directions for improvement and future research.

II. RELATED WORK

In the domain of emotion recognition, a notable approach is presented in the paper "Multimodal Speech Emotion Recognition Using Audio and Text" [8]. This method utilizes deep learning

techniques, specifically dual recurrent neural networks (RNNs), to integrate both audio signals and textual data for enhanced emotion recognition. By leveraging these two types of data, the approach achieves a more comprehensive analysis of emotional expressions. The paper reports accuracies ranging from 68.8% to 71.8% for classifying emotions into categories such as angry, happy, sad, and neutral.

Another interesting approach is presented in the paper "Multi-Modal Emotion Recognition on IEMOCAP with Neural Networks" [9]. The authors created specific models for each data type, employing architectures like LSTM, CNN, and MLP, and combined these at the final stage to enhance detection accuracy. Their innovative approach ensures flexibility and efficient retraining if a modality is absent, achieving an accuracy of 71%.

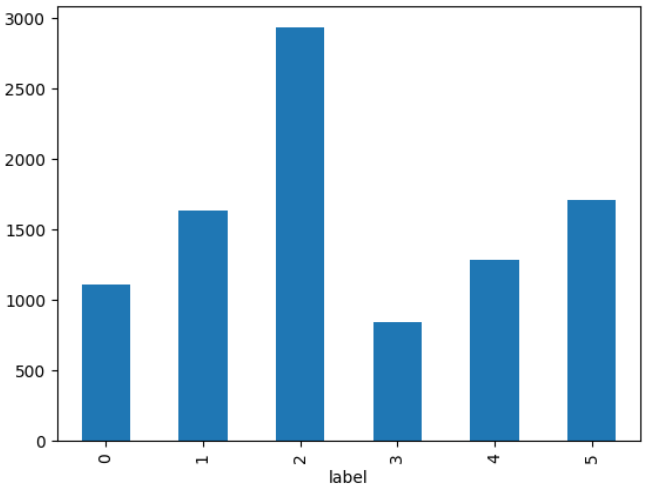
Last but not least, another deep learning approach [10] is presented by Xue Zhang, Mingjiang Wang, and Xing-Da Guo, which attained a performance metric of 68%. This model preprocesses the three modes of speech, video, and text from the IEMOCAP dataset, uses deep learning neural networks to extract emotional features, and performs information fusion at the feature layer to enhance emotion recognition accuracy.

III. DATASET

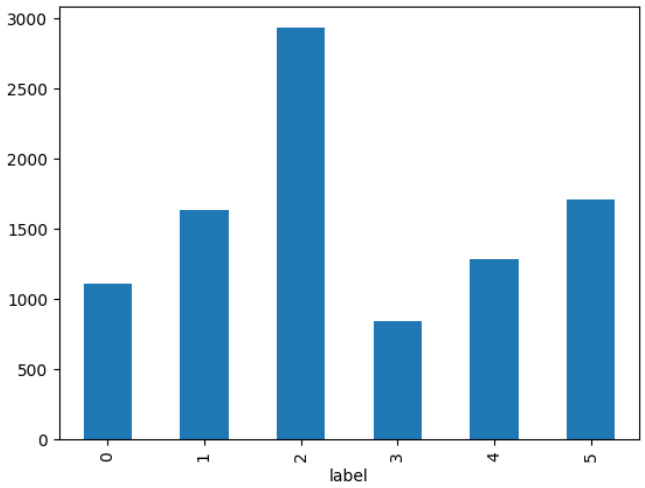
The IEMOCAP dataset is a widely used resource for emotion recognition research, consisting of approximately 12 hours of audiovisual data. This dataset includes recordings from five sessions, each featuring different pairs of male and female actors. The recordings comprise both scripted dialogues and improvisational interactions, capturing a naturalistic range of emotional expressions. The dataset is labeled with 9 categorical attributes: anger, happiness, excitement, sadness, frustration, fear, surprise, other and neutral state. It provides multimodal inputs, including audio, video, and text transcriptions, making it a comprehensive dataset for studying emotions in various contexts. The rich diversity of actors, sessions, and emotional expressions in IEMOCAP makes it an invaluable resource for advancing the development of emotion recognition models.

For preprocessing the IEMOCAP dataset, we first removed the label 'other' as it lacks a meaningful emotional context for our recognition tasks. Following trends observed in other studies, we decided to reduce the number of emotion labels. Specifically, we merged 'sad' and 'frustrated' into a single category, and combined 'happy' with 'excited,' as these pairs are closely related in our perception of emotions. This resulted in a total of six emotions: anger labeled as 0, happy/excited as 1, sad/frustrated as 2, fear as 3, surprise as 4, and neutral as 5. Then, we saw an imbalance in the dataset and decided to apply oversampling techniques to the 'fear' and 'surprise' labels, which were underrepresented. This preprocessing step ensured a more balanced distribution of emotions.

Distribution of emotions after preprocessing speech dataset



Distribution of emotions after preprocessing text dataset



Specifically for the text data preprocessing we follow the next steps. First, we removed the non-alphanumeric characters, this step let us preserve only the important information. Also, we maintained specific punctuation marks in the text (“.”,”?”,”!”), since we observed that they improved the accuracy of the models. Lemmatization is the process of converting words to their base form (“run” from “running”). Stopwords were removed since they occur frequently in a language and usually don't carry important meaning for analysis (like “the”,”and”). Moreover, we use TF-IDF

vectorizer (Term Frequency-Inverse Document Frequency) which is a numerical statistic used to indicate how important a word is to a document in a collection or corpus. The TF (Term Frequency) measures how frequently a term occurs in a document, while IDF (Inverse Document Frequency) measures the importance of the term across a set of documents. It's calculated by taking the logarithm of the number of documents in the corpus divided by the number of documents where the specific term appears.

In our analysis of feature extraction for speech recognition, we decided not to use the Mel-Frequency Cepstral Coefficients (MFCC) feature extraction method. Instead, we selected features based on the slides from our course. The primary features we chose are the mean and standard deviation of the width of the signal, energy, and pitch. Segment feature statistics such as the mean value and standard deviation are crucial for capturing temporal changes in short-term feature sequences. For pitch calculation, we utilized the Modified Autocorrelation Function (MACF) method described in "New methods of pitch extraction"[12]. Moreover, many implementations use root mean square (RMS) in order to calculate the energy for speech signal analysis. In our opinion this is a valid approach as the energy of a speech signal can be related to its loudness. Finally, we found some implementations of speech recognition to utilize a feature called silence[13]. Silence is a part of human communication and can provide clues about emotional states. For instance, pauses or silences in speech can indicate hesitation, thoughtfulness, or even emotional intensity. The study found that periods of silence in speech are more indicative of how intense the speaker's emotions are.

IV. MACHINE LEARNING MODELS

For our project we employed four machine learning models: Multinomial Naive Bayes, Random Forest, XGBoost, and Logistic Regression

Multinomial Naïve bayes employs the frequencies of words in a document as its features or predictors for classification. It is predominantly applied for classifying text. It assesses the probability of each category for a given text and selects the category with the highest probability as output. The presence or absence of a feature is considered independent of any other's feature absence or presence. By applying Bayes theorem, this "naïve" assumption of feature independence simplifies probability calculations, enhancing the algorithm's computation efficiency.

Logistic Regression is a statistical method used for binary classification. It models the probability that a given input belongs to a particular category. Logistic Regression works by fitting a logistic curve to the data and using the sigmoid function to estimate probabilities, which are then mapped to the closest class. This technique is widely used for predictive analysis to determine outcomes that have two possible states. In this work, the model predicts the resulting class based on the highest probability of 6 classifiers. The classifiers have been trained for each specific class.

Random forest is an ensemble method that combines a multitude of decision trees to make predictions. In this approach, each tree within the random forest contributes its prediction, with the most frequently predicted class becoming the final output of the model. By pooling the predictions of several independent models (trees), the collective decision is typically more accurate than that of any single tree within the ensemble. In the Random Forest algorithm, each decision tree is built using a random selection of data points and features. This means that for a dataset containing 'k' entries, 'n' random instances and 'm' features are chosen to form the foundation of each individual tree. Individual decision trees are constructed for each sample and the final output is considered based on Majority Voting for classification.

Extreme Gradient Boosting, commonly known as XGB, is using parallel processing for predicting classes effectively. It aggegrates weak learners to boost model's accuracy. XGB implements a better approach than Random forest, which trains each specific tree independently to the the others. It attempts to fix the weaknesses of the independent learners by fixing the ones that made mistakes(Forward stagewise additive modeling). The model's final prediction is calculated by linearly combining the outputs of each individual learner, weighted according to their contribution.

Ensemble methods combine the outputs of several machine learning problems to improve performance. The variety of the models can contribute to the minimization of errors and generalisation. More specifically, we utilised soft voting, a method that averages the probabilities to make improved predictions. By averaging, each classifier can fix the weakeness of the others because if one is not confident about a class, the

combined score from other classfiers can lead to correct predictions. It is usually better than using a single classifier.

V. METRICS

The evaluation metrics are derived from Mr. Giannakopoulos's slides:

- **Confusion matrix (CM)** counts samples that belong to class i and are classified to class j

$$CM(i,j), i = 1, \ldots, N_c$$

- **Recall** is the fraction of samples correctly classified to class i

$$\mathrm{Re}(i) = \frac{CM(i,j)}{\sum_{m=1}^{N_c} CM(i,m)}$$

- **Precision** is the fraction of samples correctly classified to class i if we take into account the total number of samples that were classified to class i

$$\mathrm{Pr}(i) = \frac{CM(i,j)}{\sum_{m=1}^{N_c} CM(m,i)}$$

- **F1 score** is the harmonic mean of recall and precision

$$F1(i) = \frac{2Re(i)Pr(i)}{Pr(i) + Re(i)}$$

- **Accuracy** is the proportion of data correctly classified.
- **ROC:** In order to define the performance of the models for each class within the multiclass classification problem, we use ROC curve. The ROC curve is a graphical representation that evaluates the performance of a binary classifier by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. For multiclass classification, this can be extended using the one-vs-rest (OvR) approach. In this method, each class is treated as the positive class, while the remaining classes are grouped together as the negative class. This approach ensures a comprehensive evaluation of the classifier's ability to distinguish between each class in the multiclass problem.

VI. EXPERIMENTS

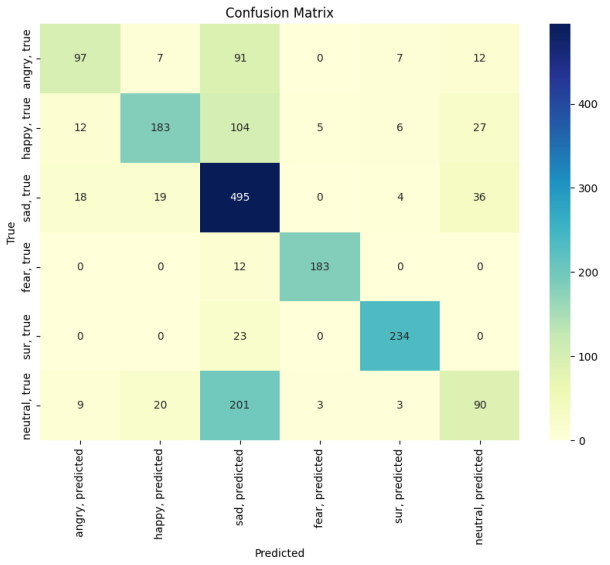
In our work, we explored three distinct experimental conditions to evaluate the performance of our classifiers. This experimental design allows us to understand the information contribution of each modality separately and examine how their concatenation affects the overall model performance.

- First condition: utilized speech feature vectors
- Second condition: utilized text feature vectors
- Third condition: combined feature vectors from both speech and text modalities

VII. RESULTS

Text (only) - accuracy	
Random Forest	61%
XGBoost	54%
Logistic Regression	61%
Multinomial NB	58%

Random Forest confusion matrix (text)



ROC curve Random forest (text)

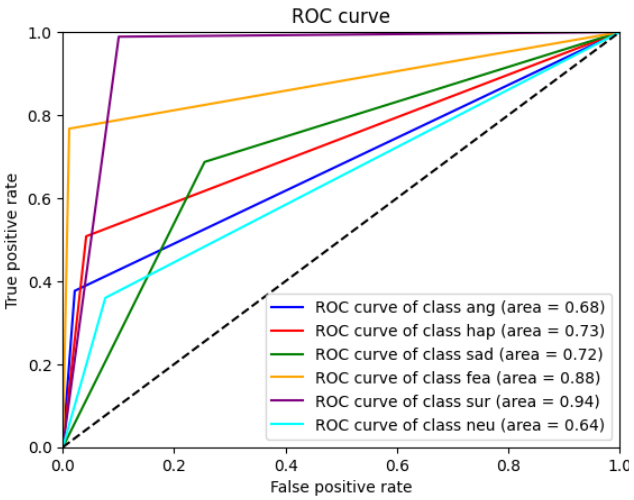


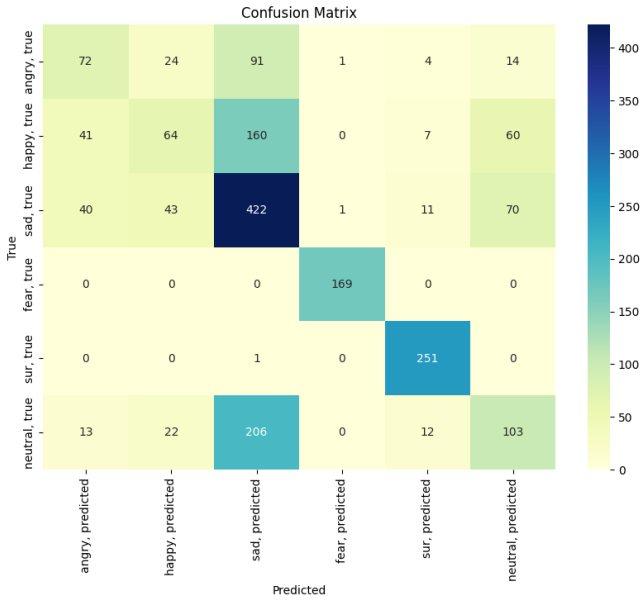
Table Random forest (text)

	precision	recall	f1	support
0	0.7	0.38	0.49	223
1	0.7	0.51	0.59	307
2	0.56	0.69	0.61	601
3	0.86	0.77	0.81	159
4	0.62	0.99	0.76	269
5	0.51	0.36	0.42	342

Text data: We observe that the "sad" and "neutral" sentiments are the most difficult to predict. Tf-idf vectors played a significant role in the increased accuracy that we observe, as compared to audio's dataset scores. Those vectors represent the correlation of the features well enough to improve the best score of the audio's dataset by approximately 5%. The best classifier on this occasion was random forest, with accuracy fluctuating around 61%. The soft-voting method did not enhance the accuracy of the model at any of our tests.

Speech (only) - accuracy	
Random Forest	56%
XGBoost	55%
Logistic Regression	32%
Multinomial NB	31%

Random Forest confusion matrix (speech)



ROC curve Random forest (speech)

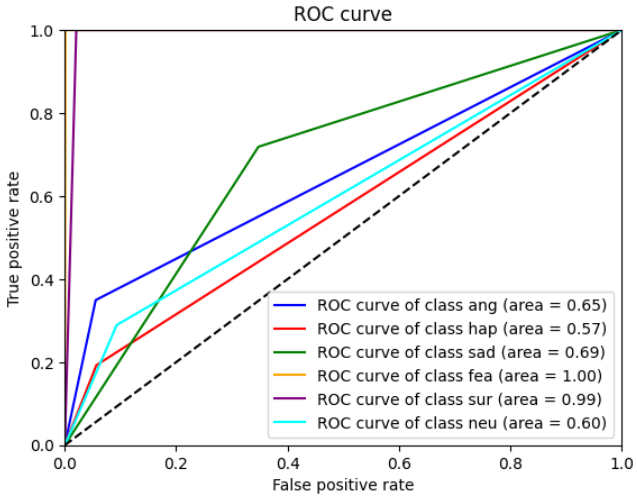


Table Random forest (speech)

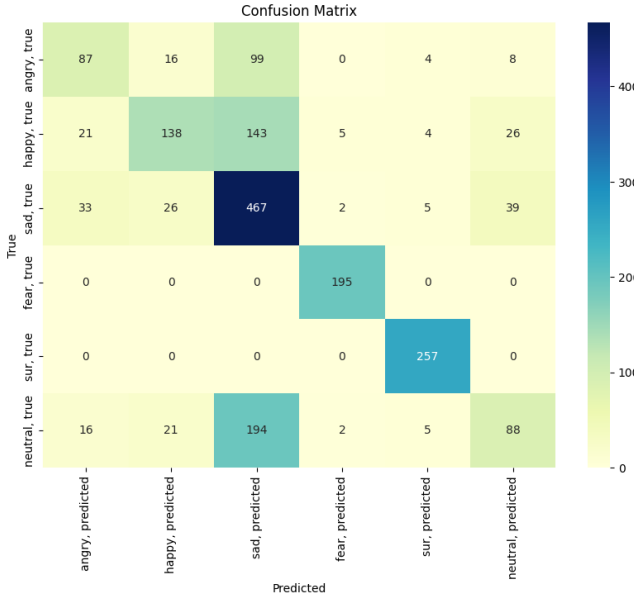
	precision	recall	f1	support
0	0.43	0.35	0.39	206
1	0.42	0.19	0.26	332
2	0.48	0.72	0.58	587
3	0.99	1	0.99	169
4	0.88	1	0.94	251
5	0.42	0.29	0.34	356

Speech data: All of the models on this experiment underperformed the text's respective ones. This was probably because the audio features

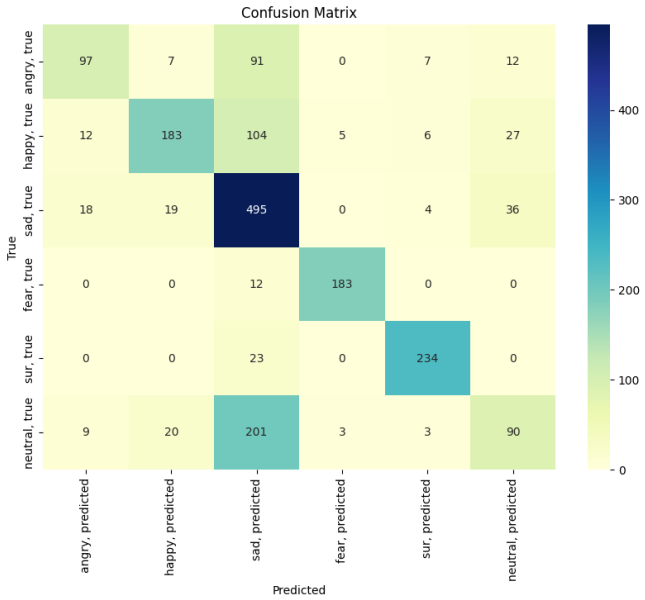
did not represent the labels accurately so that the classifier can distinguish between each class. Random forest and xgb booster achieved 56% and 55% accuracy respectively. Multinomial Naïve Bayes and multinomial logistic regression reached accuracy around 30%. The random forest struggled to distinguish between happy and sad as well as neutral and sad.

Speech + Text - accuracy	
Random Forest	64%
XGBoost	61%
Logistic Regression	60%
Multinomial NB	58%

RF confusion matrix(text+ speech)



Soft Voting RF confusion matrix (text+ speech)



Roc curve random forest (text+ speech)

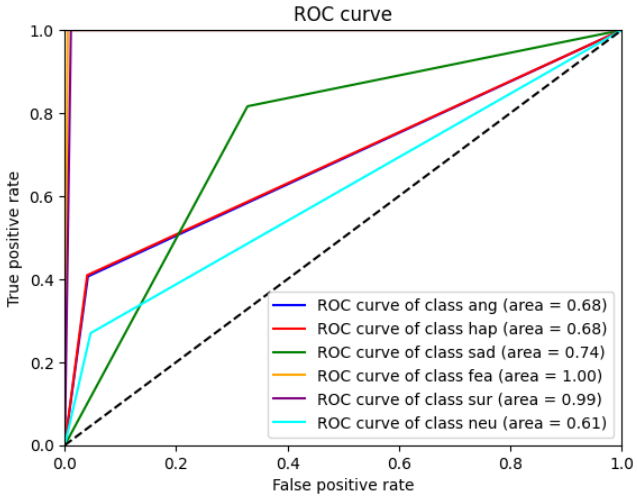


Table Random forest (text+speech)

	precision	recall	f1	support
0	0.55	0.41	0.47	214
1	0.69	0.41	0.51	337
2	0.52	0.82	0.63	572
3	0.96	1	0.98	195
4	0.93	1	0.97	257
5	0.55	0.27	0.36	326

Combined data: The combined dataset manages better results as opposed to the previous methods. The audio and text features are highly associated

with as it is indicated by the improved scores. It is worth noting that the soft voting method on all the classifiers was the more efficient than using each specific one individually. Random Forest attained 64,8% in the combined dataset while soft voting one 67,8%. By averaging classifiers, we leverage the strengths of each classfiers, resolving the weakenesses and achieving the best score of the experiment. The soft-voting helped eliminate the ambiguity observed in the individual models since the classifier could distinguish better between happy and sad as opposed to random forest method. RF predicted correctly 201 “happy” samples from the total 337 ones while soft voing found properly 229 happy samples from the total 337.

Regarding the roc curve scores, we understand that when the line diagram for one class cover greater area, this translates to better model’s accuracy. We desire the line diagram to be as close as possible to the upper left corner of the plot. We observe that lines of the 3 and 4 class tend to be closer to the upper left corner, which means that they are predicted more accurately than the else classes .

Comparison of our Best ML model to others

Model	**Accuracy**
Random Forest	65%
Approach [8]	68% – 71%
Approach [9]	71%
Approach [10]	68%

The best performing machine learning model using the combined dataset of speech and text is the Random Forest, achieving an accuracy of 65%. As shown in the table, this performance is comparable to other approaches that utilize deep learning models. This comparison indicates that our Random Forest model's performance is close to these deep learning models, highlighting its effectiveness despite the simplicity of the approach.

VIII. FUTURE WORK

To enhance the performance of our machine learning models for emotion recognition from speech and text, we plan to implement several advanced strategies. First, we will split the data into scripted and improvised segments to better understand how these different types of speech influence emotional expression and model performance. Additionally, we will employ a leave-one-session-out cross-validation approach to ensure robust evaluation by testing the model's ability to generalize across different sessions. For speech data, we intend to utilize spectrograms for feature extraction, capturing the detailed frequency components of the audio signals. For text data, we will leverage advanced word embeddings such as BERT or GloVe, which are known for their contextual understanding and representation of text.

In our final enhancement to the emotion recognition models, we plan to implement the approach described in the paper "A Deep Learning Method Using Gender-Specific Features for Emotion Recognition" [11]. This method leverages gender-specific features to enhance the accuracy of emotion recognition in speech. Initially, the speech data is classified by gender using a Multi-Layer Perceptron (MLP). Subsequently, distinct acoustic features of male and female speech are analyzed to establish optimal feature sets for each gender. Separate models, including Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory networks (BiLSTM), are then trained using these gender-specific feature sets. This approach has been shown to significantly improve the accuracy of emotion recognition compared to models that do not differentiate between genders.

IX. REFERENCES

[1] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.-M. Lee, A. KazemzadehS. Lee, U. Neumann, and S. Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information,”

[2] D. Ververidis and C. Kotropoulos, “Fast and accurate sequential floating forward feature selection with the bayes classifier applied to speech emotion recognition

[3] T. Vogt and E. Andr ´e, “Comparing feature sets for acted and sponta-neous speech in view of automatic emotion recognition,”

[4] Al Hanai, T., Ghassemi, M., & Glass, J. (2018). ‘’Depression detection using emotion artificial intelligence’’

[5] M. Dhuheir, A. Albaseer, E. Baccour, A. Erbad, M. Abdallah and M. Hamdi, "Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey,"

[6] Blumentals, E., & Salimbajevs, A. Emotion Recognition in Real-World Support Call Center Data for Latvian Language.

[7] Lawpanom, R., Songpan, W., & Kaewyotha, J. (2024). Advancing Facial Expression Recognition in Online Learning Education Using a Homogeneous Ensemble Convolutional Neural Network Approach.

[8] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, 'Multimodal Speech Emotion Recognition Using Audio and Text'

[9] Tripathi, S., Tripathi, S., & Beigi, H. Multi-Modal Emotion Recognition on IEMOCAP with Neural Networks

[10] X. Zhang, M. Wang, and X.-D. Guo, "Multi-modal Emotion Recognition Based on Deep Learning in Speech, Video and Text

[11] L i, Y., Zhang, Y.T., Ng, G.W., Leau, Y.B., & Yan, H. (2023). ''A Deep Learning Method Using Gender-Specific Features for Emotion Recognition''

[12] M. Sondhi, "New methods of pitch extraction"

[13] Bagus Tris Atmaja1, Masato Akagi 'The effect of silence feature in dimensional speech emotion recogniton''