Multimodal Performance Analysis during job interviews

by

Amina Aman

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

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

May 2023

Author ..... Amina Aman
Department of Computer Science
April 29, 2023

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

Accepted by .....
Vassilios D. Tourassis
Dean, School of Engineering and Digital Sciences

Multimodal Performance Analysis during job interviews

by

Amina Aman

Submitted to the Department of Computer Science
on April 29, 2023, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Emotion recognition based on multimodal data has become an important research topic with a wide range of applications, including online interviews. The study of respondents' performance through the analysis of multiple modes of data is essential for a deep understanding of their emotions and communication patterns. To solve this problem, this thesis proposes a new method of analyzing multimodal interviews that uses deep learning techniques to extract meaningful information from various sources, such as video, audio, and textual data. The proposed approach uses late fusion to integrate information from different sources and generate an overall summary of the interviews. The effectiveness of the proposed method is evaluated on the whole MIT interview dataset, which includes 138 mock job interviews conducted with MIT undergraduates. The experimental results demonstrate that our framework can efficiently analyze multimodal data to produce promising results. The proposed approach identifies and captures critical aspects of communication, such as tone, facial expressions, and language use, which can provide valuable information to interviewers to improve the overall interview process. This research has implications for improving understanding of communication patterns in various contexts, including job interviews, and may have practical applications in other fields.

Thesis Supervisor: Adnan Yazici
Title: Department Chair of Computer Science

Acknowledgments

I want to sincerely thank my thesis supervisor, Professor Adnan Yazici, for his important advice and ongoing support during the whole research. For their unwavering support during this difficult academic year, I am extremely grateful to my family and friends.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The analysis of human emotional behavior is increasingly important in the field of Machine Learning Intelligence. Emotions play a significant role in communication, and understanding them can lead to better decision-making in various applications. For instance, analyzing social relationships can help in improving human-computer interaction, while analyzing depression can provide early intervention for people who are suffering from mental health issues. Additionally, analyzing job candidates' emotional behavior can lead to more informed hiring decisions.

With the pandemic, video-based communication has become the norm. Therefore, researchers are now using video-based communication modes to study human emotions. Facial expressions, words, tone, and gestures are some of the different channels through which humans express their emotions. Being able to see and hear these expressions makes it easier to understand each other's emotions.

Analyzing emotions through multiple channels can provide more reliable results than relying on a single channel. Multimodal Interview Analysis has primarily focused on identifying basic emotions or personality traits through the use of deep learning techniques. For instance, studies like [15], [26], and [27] have focused on predicting labels such as extraversion, agreeableness, conscientiousness, neuroticism, and openness, which are considered to be basic personality traits. These models leverage various data modalities such as video, audio, and text to extract features and develop an accurate representation of the candidate's behavior. However, studies have

shown that focusing on personality traits alone may not be enough to make informed hiring decisions. To address this, Naim et al. [1] developed an annotated dataset consisting of job interview videos taken from MIT undergraduates who were seeking internships. This dataset includes high-level behavioral dimensions such as warmth, presence, competence, and content labels, which can provide more valuable information for making informed hiring decisions. The authors proposed machine learning techniques to analyze videos and predict the emotions, interests, and competence of candidates.

To further enhance the analysis of verbal and non-verbal behavior cues, this study proposes and evaluates deep learning methods for multimodal interview analysis. The proposed approach captures high-level behavioral dimensions and critical aspects of communication, such as tone, facial expressions, and language use. It provides valuable information to interviewers to improve the overall interview process. The study explores the use of Convolutional Neural Networks (CNNs) for visual data, Long Short-Term Memory Networks (LSTM) for the text part, and Feed-Forward Neural Networks (FNN) for audio data from interviews independently. All results, including emotions, interest, and competence of the candidate, are concatenated in a vector in the late fusion stage to make the final conclusion if the candidate is "Recommended Hiring" or "Not Recommended" using Feed-Forward Neural Network (FNN). Additionally, ensembling technique using Random Forest Regressor is presented as another way of getting final hiring decision allowing us to see success rate of the candidate.

The study's significance lies in its implications for enhancing understanding of communication patterns in various contexts, including job interviews, and may have practical applications in other fields. The effectiveness of the proposed approach was demonstrated on the MIT interview dataset, which contains 138 interview sessions with 69 undergraduates who were seeking internships. According to experimental results, the suggested framework produced encouraging results, showing 92.3% accurate prediction using FNN and 94% using Random Forest Regressor on final classification. This study's findings could lead to more accurate and informed hiring decisions and could also be applied to other contexts where the analysis of human emotional be-

havior is relevant.

In this paper, we aim to propose multimodal interview analysis system that has the potential to provide invaluable insights for hiring recommendations by utilizing deep learning techniques. In order to achieve this goal, we have structured our paper into several sections. In Section 2, we review the related works and existing literature on this topic, providing a comprehensive overview of the current state of the art. In Section 3, we describe our methodology and the steps we took to implement our approach. Section 4 presents the results of our experiments and the evaluation of our approach, with a detailed analysis of the outcomes. Finally, in the last chapter, we summarize our findings and discuss the implications of our research, highlighting both the strengths and limitations of our approach.

Chapter 2

Related works

2.1 Multimodal Video Analysis

As mentioned earlier, multimodal video analysis has received significant research attention. Years of thorough study have demonstrated that multimodal systems are more effective than unimodal systems in understanding the emotion of a speaker. Textual, acoustic, and visual modalities are frequently combined to extract information from communication in an efficient manner. This is how humans naturally communicate and express their feelings and sentiments [8].

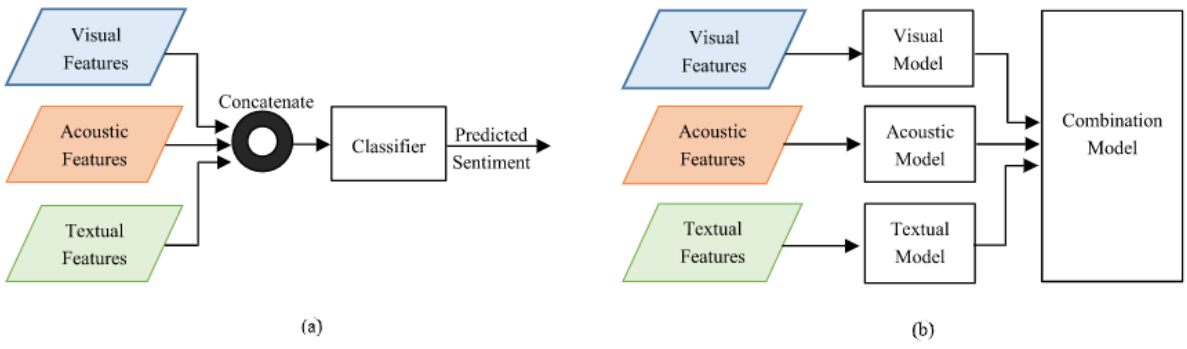

Figure 2-1: Traditional Fusion Techniques

The early fusion-based prediction models (Fig 2.1a) could be as straightforward as Hidden Conditional Random Fields (HCRFs), Support Vector Machines (SVMs),

or Hidden Markov Models (HMMs). Recurrent Neural Networks, particularly Long-Short Term Memory, have been employed for sequence modeling later, following the advancements of deep learning. As in [5], to represent the link between emotions, Dai et al. propose a modality-transferable network with cross-modality emotion embeddings. Three LSTM networks, one emotion embedding mapping module, and one modality fusion module comprise the architecture of their proposed multi-modal emotion identification model.

Whereas, by combining audio and visual information at the model level, an optimum multimodal emotion detection model using CNN-LSTM is constructed in the paper [4]. On the SAVEE [9] and RAVDESS [10] datasets, the suggested models demonstrate high predicted accuracies of 99% and 86%, respectively.

The approach of late fusion involves developing distinct models for each modality and subsequently merging their outcomes through techniques such as averaging, weighted sum, majority voting, concatenation of prediction from different modalities to one vector, or deep neural networks, as illustrated in Fig. 2.1b. The late fusion approach combines many modalities at the level of prediction [11]. Ding et al, [12], offer two model variations based on various experimental conditions, a multi-level late-fusion learning framework with residual connections, and a more sensible experimental data-set split. Also, a good example of late fusion represented in [14], offers a completely new deep neural network (DNN) that fuses audio, video, and text modalities to recognize emotions across many media.

2.2 Multimodal Interview Analysis

In the context of interviews, multimodal analysis can provide a richer and more nuanced understanding of the interviewee's behavior, emotions, and cognitive processes.

Multimodal Interview Analysis approaches aim to identify various traits and characteristics of job candidates from multiple modalities, including audio, video, and text. Several deep learning models have been proposed to predict basic emotions or basic personality traits, such as [15], [26], and [27]. Li et al. [16] predict the Big

Reference	Year	Dataset	Classification Algorithms	Recognized labels	Accuracy avg
[16] Li et al.	2020	First Impression v2 dataset	Deep Classfication-Regression Network (CR-Net)	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism(OCEAN), job interview recommendation.	0.9188
[26] Agrawal et al.	2018	First Impression v2 dataset	Multi-task deep neural network (MTDNN)	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism(OCEAN), job interview recommendation.	0.9134
[27] Kaya et al.	2019	First Impression v2 dataset	Extreme Learning Machine (ELM) classifiers	Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism(OCEAN), job interview recommendation.	0.9170

Table 2.1: Papers using First Impression v2 dataset

Five personality traits and further assists on job interview recommendation by using the CR-Net architecture that uses ResNet-34 as its backbone network to analyze multimodal data such as audio, video, and text. The final prediction is obtained by fusing features from different modalities and using ETR. In [26] multi-task deep neural network(MTDNN) is constructed to predict personality traits like Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (commonly abbreviated as OCEAN) and job interview scores of individuals based on the scene, audio, and facial features. The same the First Impressions v2 dataset was applied by Kaya et al. [27] using extreme learning machines (ELM) and feature extraction.

Naim et al. [1] collected, utilized, and trained Lasso and SVR models on the MIT While previous multimodal interview analysis studies have primarily focused on identifying basic emotions or personality traits, the MIT Interview dataset contains more valuable information for making informed hiring decisions. This includes high-level behavioral dimensions such as warmth (e.g. friendliness, smiling), presence (e.g. engagement, excitement, focused), competence (e.g. speaking rate), and content (e.g. structured answers) labels. The ground truth labels were obtained by averaging the ratings of nine Amazon Turk employees, who manually gave the labels to these ratings.

In [2], Agrawal et al. provided a multimodal analytical framework that evaluates the interviewee in terms of engagement, speaking rate, eye contact, etc. in order to offer feedback for predetermined labels. They employ classification, with the numbers 1 to 7 denoting the degree of performance, as opposed to [1] which used regression

Reference	Year	Dataset	Classification Algorithms	Recognized labels	Accuracy avg
[1] Naim et al.	2018	MIT Interview Dataset	Lasso, SVR	Overall, Recommend Hiring, Engagement, Excitement, Eye Contact, Smile, Friendliness, Speaking Rate, No Fillers, Paused, Authentic, Calm, Focused, Structured Answers, Not Awkward	AUC avg 0.80
[2] Agrawal et al.	2020	MIT Interview Dataset	Random Forest Classifier, SVC, Multitask Lasso, MLP	Eye contact, Speaking Rate, Engaged, Pauses, Calmness, Not Stressed, Focused, Authentic, Not Awkward	Avg Accuracy 0.74
[16] Chopra et al.	2020	MIT Interview Dataset	SVR, KNN, Decision Tree	Friendly, Engaged, Excited, Speaking Rate, Calm	AUC avg 0.72

Table 2.2: Researches done on MIT Interview Dataset.

as an evaluation measure. As classifiers, four distinct algorithms—Random Forest, SVC, Multitask Lasso, and MLP—were employed.

One more research on Interview Data Analysis is [16] by Chopra et al. They examine the MIT Interview dataset to determine the relationship between certain prosodic variables, such as pitch, intensity, and others, and the prospect of receiving a positive evaluation during the interview. They used three regression models - Decision tree, SVR and KNN. And revealed that the Decision tree is the best choice from all three models for prediction.

As summarized in Table 2.2, despite the potential value of information given in the MIT interview dataset, deep learning algorithms have not yet been used on it. Therefore, this thesis research aims to fill this gap by developing a deep-learning multimodal architecture that can accurately capture and classify these high-level behavioral dimensions. This architecture will combine video, audio, and text data to extract relevant features and develop an accurate representation of the candidate's behavior. The resulting model will allow recruiters to make more informed hiring decisions and identify the best candidates for the job.

Chapter 3

Methodology

3.1 Dataset

To verify our experiment, we employed the MIT interview dataset [1], which comprises recordings of 138 simulated job interviews conducted with 69 participants before and after the intervention. There are 138 interview videos totaling around 10.5 hours in length, or 4.7 minutes for each interview on average. They claimed that it represented the biggest collection of video interviews performed in real-life circumstances by licensed counselors [1].

The dataset contains evaluations from Amazon Mechanical Turk Workers for each video, which are aggregated to determine the final score for each label. Additionally, the dataset includes audio files, which we analyze for audio processing and has transcripts from each interview for text classification. This dataset uses a classification system where performance levels are denoted by numbers 1 to 7. A score of 1 indicates a very poor performance for any label, while a score of 7 is regarded as outstanding.

Five distinct questions, all of which were suggested by MIT Career Services, were posed by the counselor to each interviewee. The candidates for the interview received no job description. The interviewee's social and behavioral abilities were evaluated through the use of the five questions which are given in Table 1.

Figure 3-1: MIT Interview Dataset

Q1	So please tell me about yourself.
Q2	Tell me about a time when you demonstrated leadership
Q3	Tell me about a time when you were working with a team and faced a challenge.
Q4	What is one of your weaknesses and how do you plan to overcome it?
Q5	Now, why do you think we should hire you?

Table 3.1: List of Interview questions

3.2 Proposed Model

The Proposed Model of Multimodal Interview Analysis is an innovative approach to analyzing interviews that incorporates multiple modalities, including facial expressions, audio signals, and text-based responses. The model utilizes Convolutional Neural Networks (CNNs) for facial modality analysis, Feedforward Neural Networks (FNNs) for audio modality analysis, and Long Short-Term Memory (LSTM) networks for text modality analysis. The results from these modalities are then concatenated

and fed into a Feedforward Neural Network (FNN) for final fusion and classification.

The CNN component of the model is responsible for analyzing facial expressions and identifying emotions displayed by the candidate during the interview. The FNN component processes the audio signals from the interview, including tone, pauses, and rhythm. The LSTM component analyzes the candidate's text-based responses to interview questions, including sequence of the words and overall content. By combining the results from these modalities, the model can gain a more complete understanding of the candidate's overall performance and suitability for the job.

The final fusion and classification component of the model uses a Feedforward Neural Network (FNN) to classify the candidate as either recommended or not recommended for the job. The FNN takes the concatenated results from the CNN, FNN, and LSTM modalities and feeds them into a fully connected layer for final classification. The output of the FNN represents the model's recommendation for hiring the candidate.

Overall, our aim is to gain valuable information from each modality and concatenate everything to get one final result. Each modality and final classification looks as in Figure 2 and will be separately explained in each of subsection of this chapter.

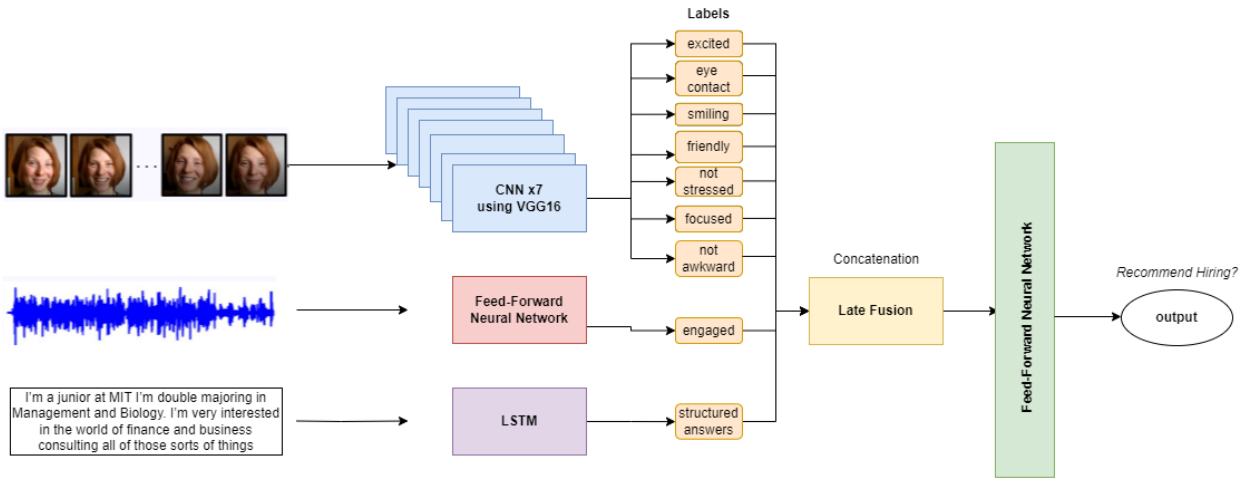

Figure 3-2: Proposed model architecture

3.2.1 Visual Modality

Data preprocessing

The first step in our pipeline is to extract frames from the video at regular intervals. We use the get_frames() function to extract a frame every 3 seconds. Next, we use the Haar Cascade algorithm [17] to detect faces in each frame. This algorithm is trained to detect facial features such as the eyes, nose, and mouth, and can accurately locate the face in an image. Once we have located the face, we use the save_cropped() function to crop the face and save it for further processing. The example of cropped face from the dataset can be observed in Figure 3-3.

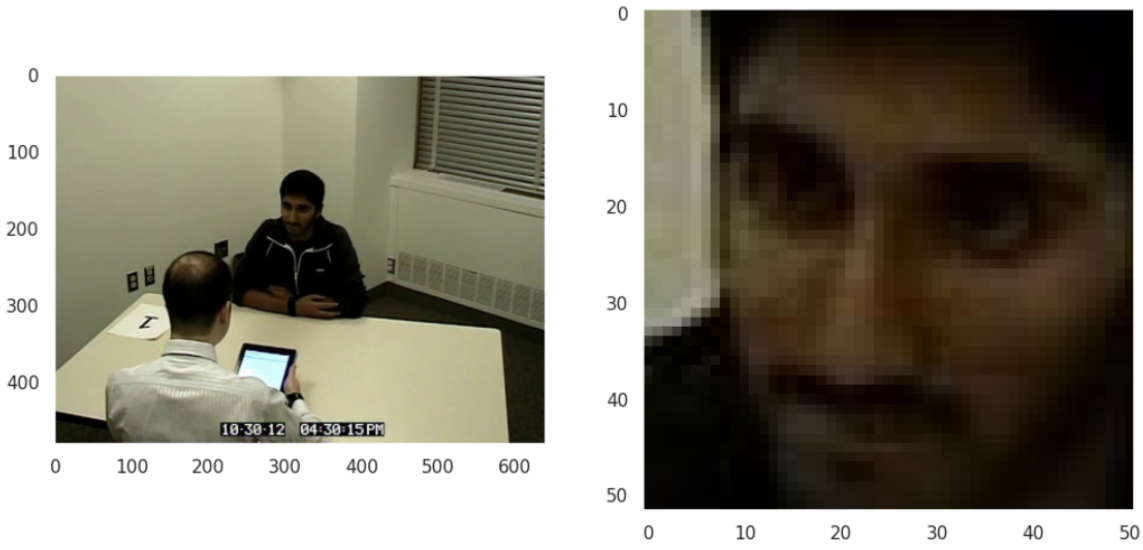

Figure 3-3: Detected and cut face from Video frame

To improve the generalization and robustness of our deep learning model, we apply several image augmentations to the cropped face images. We use OpenCV's image processing functions to perform various augmentations on the cropped face images. These augmentations include grayscale conversion, flipping, rotation, padding, and normalization. Grayscale conversion reduces the dimensionality of the data and simplifies the input for the model. Flipping and rotation provide additional variations of the input data, which can improve the model's ability to recognize faces in different orientations. Padding is used to standardize the size of the face images, while

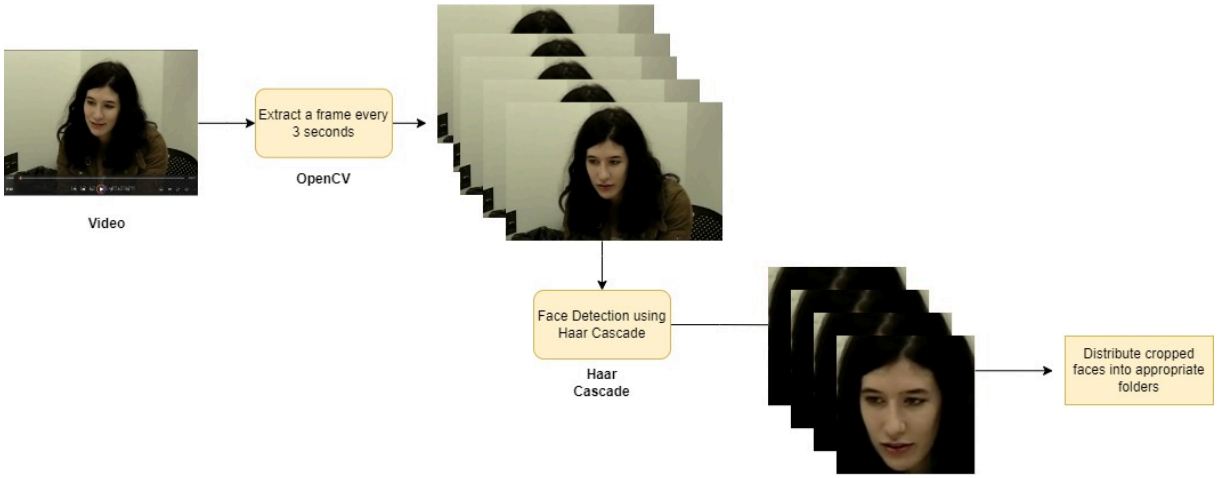

Figure 3-4: Preprocessing Video

normalization ensures that the pixel values are within a certain range.

For the face part, we designated emotion labels such as 'friendly', 'focused', 'awkward', 'eye contact', 'excited', 'stressed', and 'smiling', which are easy to determine from visual content. Extracted cropped faces were distributed to the appropriate directories like smiling/not smiling, excited/not excited, friendly/not friendly, focused/not focused and etc. manually.

Modelling

In order to identify emotions from images, we used CNN architecture. As some of these emotions are synonyms to each other, there are separate CNN models for each label. For that, we use transfer learning using the Visual Geometry Group-16 (VGG-16) classification model [18], which was pre-trained on the ImageNet dataset and fine-tuned for emotion classification.

The VGG16 architecture consists 13 convolutional layers, 5 max-pooling layers, and 3 fully connected layers as shown in Figure 3-5. The convolutional layers are responsible for extracting features from the input image, while the fully connected layers classify the image based on these features. The first layer in the network takes in the raw pixel values of the input image, and subsequent layers extract increasingly complex features. The max-pooling layers reduce the spatial dimensions of the feature

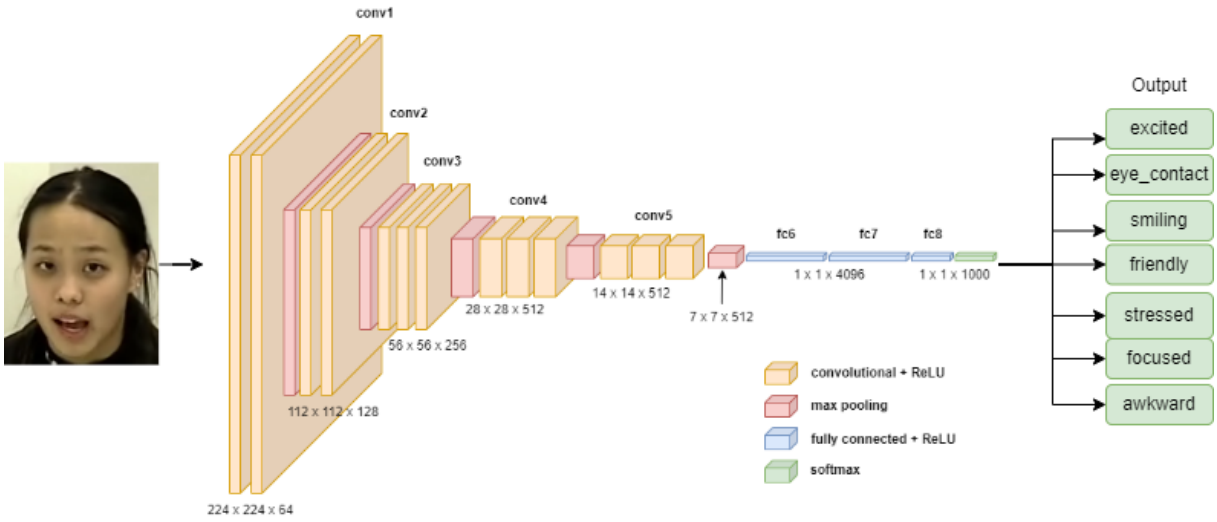

Figure 3-5: VGG16 model architecture [18]

maps, which helps to reduce the number of parameters in the model.

For transfer learning, we use the pre-trained weights of the VGG16 model, which were trained on the large-scale ImageNet dataset. By using pre-trained weights, we can leverage the knowledge gained from ImageNet to improve the performance of our model on a smaller dataset. We replace the final fully connected layer of the VGG16 model with a new fully connected layer that has the same number of output classes as our target dataset. We then freeze the weights of the convolutional layers and only train the weights of the final fully connected layer.

To further improve the performance of our transfer learning model, we employ data augmentation techniques such as flipping, rotation, and scaling. These techniques help to artificially increase the size of our dataset, which can help to prevent overfitting and improve the generalization performance of the model.

In summary, our research utilizes the VGG16 model as a transfer learning model for image recognition tasks. We replace the final fully connected layer and fine-tune the weights of the model on our target dataset. We also use data augmentation techniques to further improve the performance of our model. Through experimental evaluation, we show that our transfer learning approach using VGG16 can achieve high performance on benchmark datasets while requiring less training time and fewer computational resources than training a model from scratch.

3.2.2 Audio Modality

Data preprocessing

In order to accurately describe the interviewee's speaking style, prosodic features are crucial. In MIT Interview Dataset [1] interviews are in style of dialogue between interviewer and interviewee. Interviewees' responses were extracted and audio were cut into small excerpts according to each sentence of the candidate. In order to analyze the speaking style and emotions, prosodic variables including frequency, pitch information, tone, intensity, spectral energy, spectral centroid, zero-crossing rate and etc. are regarded to be fundamental [2].

For feature extraction process, we use [25] and Librosa audio library. It provides a wide range of functions for tasks such as reading audio files, generating spectrograms and chromagrams, computing Mel-Frequency Cepstral Coefficients (MFCCs), and extracting beat. In order to identify tone being engaged or not, the important audio features are "pause_number", "avg_pause_length", "rhythm_mean" and "power_mean".

- The pause number is identified by counting the number of consecutive zero-crossings in the audio signal, which is a common way to detect silence or pauses in audio.

- The average pause length is then calculated by taking the total length of all the pauses detected and dividing it by the number of pauses. This gives an average pause length in seconds.

- The rhythm mean is the mean value of the inter-pause intervals, which is calculated by subtracting the end time of each pause from the start time of the next pause, and taking the mean of these values. This gives an estimate of the average rhythm or tempo of the speech.

- Finally, the power mean is the mean value of the power spectrum of the audio signal. This gives an estimate of the overall loudness or energy of the speech.

It is calculated using a Fourier transform, which decomposes the audio signal into its constituent frequencies and their amplitudes.

Extracted features are written down in the tabular format (Figure 3-6) and the target column to predict is taken as the candidate is "engaged" or "not engaged".

	A	B	C	D	E	F
1	id	pause_number	avg_pause_length	rhythm_mean	power_mean	label
2	p1_s12	7	1.578542274	0.4864430272	0.1012624038	1
3	p2_s2	12	0.587755102	0.4291078535	0.1013849548	1
4	p2_s6	8	1.133061224	0.4872235956	0.1025033748	0
5	p2_s37	6	6.343401361	0.5907422019	0.1041162456	0
6	p2_s25	17	0.7083793517	0.4862344352	0.09545468861	1
7	p30_s14	0	0	0.5556991935	0.05333589916	1
8	p1_s2	13	0.7736263736	0.4848572247	0.07850269153	1
9	p30_s16	5	1.196798186	0.4612550612	0.07967601097	0
10	p1_s7	8	1.005714286	0.4721433986	0.07625260168	1
11	p30_s10	0	0	0.627123818	0.0659661322	0
12	p1_s11	11	1.185009276	0.4817593399	0.1009919675	1

Figure 3-6: Prepared table for audio modality

Modelling

TabularModel from PyTorch was chosen to handle data frame tabular data. Deep learning tabular models in PyTorch typically utilize Feed-Forward Neural Network (FNN) architectures that consist of several layers of fully connected neurons with activation functions in between; as an input, data is represented as tensors with its predicted output as another tensor; during training the model learns to adjust weights and biases using backpropagation and gradient descent in order to minimize loss function which measures the difference between predicted output and actual output [19].

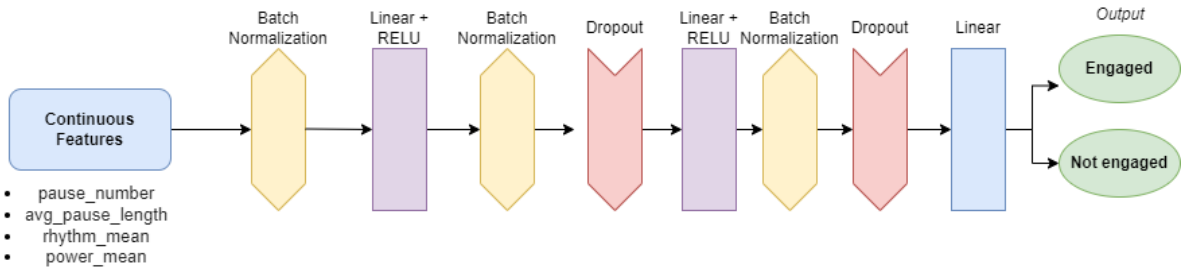

Figure 3-7: FNN for Audio Modality

The first layer is a fully connected linear layer with input size of 4 (corresponding to the number of continuous variables) and an output size of 200. This is followed by a ReLU activation function and batch normalization layer. The next layer is another linear layer with an input size of 200 and an output size of 100, followed by a ReLU activation function and batch normalization layer. Finally, there is a third linear layer with an input size of 100 and an output size of 2, which corresponds to the number of output classes (corresponding to two class labels in the dataset: Engaged or Not engaged); Overall, the model takes in 4 continuous variables as input, applies linear transformations and ReLU activation functions, and produces a binary classification output.

3.2.3 Lexical modality

Data preprocessing

As mentioned before, in the MIT interview dataset [1], interviews are given in a dialogue format between the interviewer and the candidate. The transcript is provided in annotated form with the beginning and finish of each interviewee's response to help distinguish between the interviewer's and interviewee's speech. Accordingly, the interviewers' part was removed from the text in order to analyze the candidates' responses. We divide the full text into sentences, remove punctuation, lower the words and tokenize each sentence.

Modelling

LSTM, a type of recurrent neural network (RNN), is a powerful tool for sequence data processing and is especially useful for text sentiment analysis. Compared to a standard neural network, RNNs can handle sequence-changing data, where the meaning of a word can change based on the context that comes before it.

LSTM is a specialized type of RNN that can effectively solve the problems of gradient disappearance and explosion that occur during long sequence training. Its ability to retain information for a long time makes it ideal for processing longer sequences, which is often necessary for accurate sentiment analysis [4].

Therefore, we can use an LSTM-based model to perform text sentiment analysis, which is one of the most successful RNN variants for this task. By leveraging the power of LSTM, we can improve the accuracy and effectiveness of the sentiment analysis model. Specifically, this consisted of an embedding layer that converts each token in the input sequence into a dense vector representation of fixed size followed by Long Short-Term Memory (LSTM) cells to process input sequences and identify sequential dependencies between words that capture the meaning of the text. Each LSTM layer had 256 hidden units; to prevent overfitting there is also a dropout layer with a dropout rate of 0.5 between every layer of Long Short-Term Memory cells.

The LSTM layer consists of three gates as represented in Figure 3-9: the input gate, the forget gate, and the output gate. These gates control the flow of information through the LSTM cell, allowing it to selectively store or discard information based on its relevance to the task at hand. The hidden units within the LSTM layer can learn to capture the sentiment of the sentence by considering the context of each word in the sentence, including its position, context, and relationships with neighboring words.

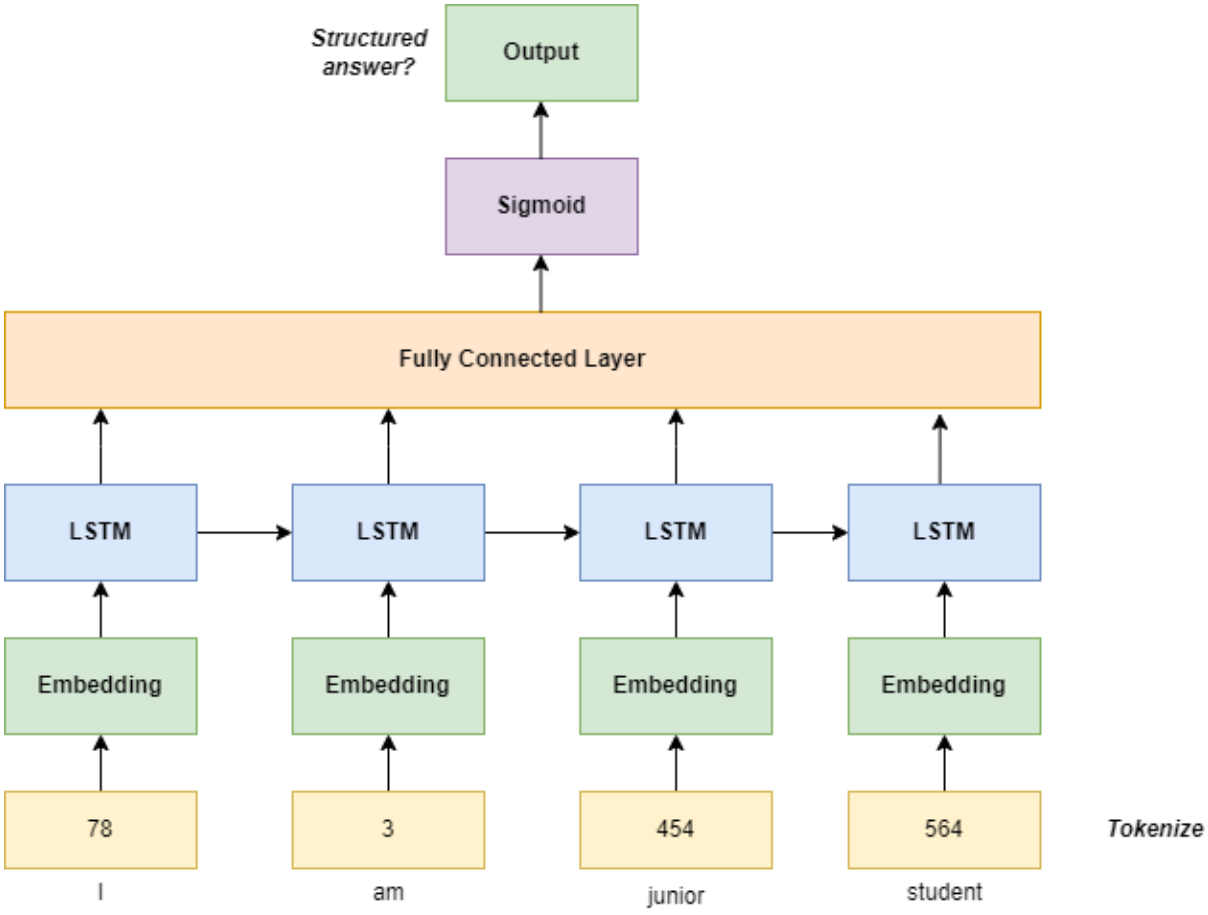

Figure 3-8: Sentiment Analysis model using LSTM for text

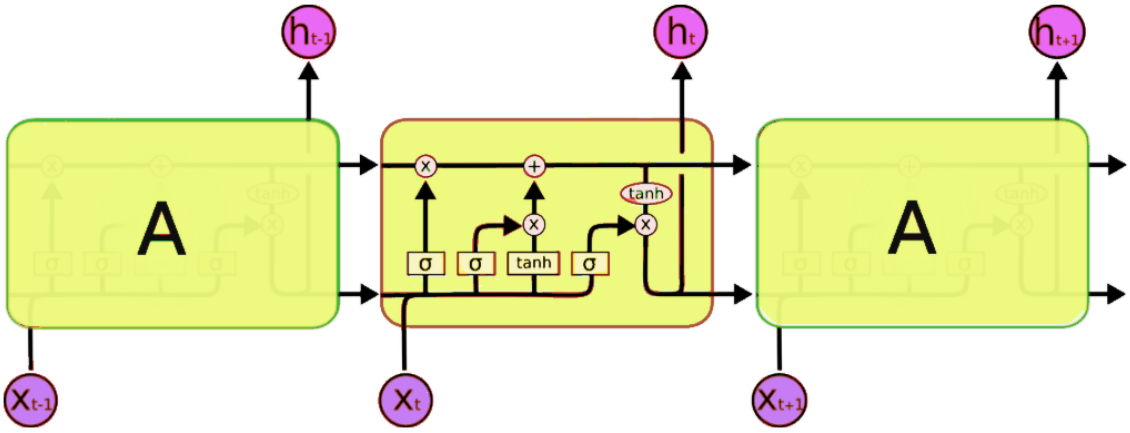

Figure 3-9: Architecture of LSTM

Using 256 hidden units in an LSTM layer can allow the model to capture more complex patterns in the data, resulting in better performance on sentiment analysis tasks. However, larger hidden units may increase the computational complexity of the model and require more training data. Therefore, it is important to balance the number of hidden units with the available training data and computational resources.

After processing an input sequence, its output from the last LSTM cell is sent through a fully connected layer with one output unit and then through a sigmoid activation function that converts its output to probabilities between 0-1; these probabilities represent the likelihood that an input text has "structured answer" or not and can be observed visually in Figure 3-8.

Overall, this model is an effective architecture for performing sentiment analysis tasks [20], and can be trained using a binary cross-entropy loss function to predict whether an input text should be considered a "structured answer" or "Not structured answer".

3.2.4 Fusion and Final Classification

As we mentioned earlier, in the fusion stage, the various modalities or features are processed independently by several models, and the results are then integrated. So, our final decision was to put all the data into a data frame, which will be a collection

of labels from visual, audio, and text models.

emotion	engagement	excited	eye_contact	smiling	friendly	paused	not_stressed	focused	speaking_rate	not_awkward	structured_answers
0	True	0.625	0.232143	0.410714	0.25	0.285714	0.464286	0.017857	0.573237	0.928571	True

Figure 3-10: Final concatenated data frame of labels

Data preprocessing

The MIT Interview dataset [1] contains evaluations from Amazon Mechanical Turk Workers for each video, which are aggregated to determine the final score for each label and given in CSV file (Figure 3-11). Since the dataset has different metrics and rates everything from 1-7, we need to divide it by 7 to get the range from 0-1. This will put our predictions to the same representation. For the "recommend_hiring" we use a threshold (> 5 = passed, not passed otherwise) for the final classification.

	recommend_hiring	engagement	excited	eye_contact	smiling	friendly	paused	not_stressed	focused	speaking_rate	not_awkward	structured_answers
0	1	1	0.720556	0.838017	0.510880	0.750683	0.828638	0.764394	0.835032	0.801502	0.782505	1
1	0	1	0.800227	0.775266	0.866025	0.928618	0.684507	0.788745	0.792833	0.859280	0.703364	1
2	0	1	0.608763	0.551121	0.693164	0.767126	0.739635	0.795155	0.760591	0.808338	0.636130	1
3	1	1	0.672437	0.956325	0.560068	0.807017	0.840904	0.845196	0.903155	0.860010	0.804263	1
4	0	1	0.664037	0.618829	0.599175	0.642196	0.718260	0.832056	0.802774	0.767408	0.756568	1
...	...	...	...	...	...	...	...	...	...	...	...	...
132	1	1	0.827355	0.895273	0.832605	0.852255	0.883572	0.852821	0.942836	1.000000	0.900193	1
133	1	1	0.829517	0.835397	0.926147	0.935611	0.823507	0.770207	0.869090	1.000000	0.834364	1
134	1	1	0.758180	0.795000	0.933372	0.932789	0.844631	0.821524	0.855455	0.857143	0.738564	1
135	1	1	0.797833	0.897180	0.721559	0.852964	0.806250	0.714136	0.936301	0.857143	0.685799	1
136	0	1	0.733949	0.841847	0.696490	0.717455	0.687750	0.848545	0.725494	0.857143	0.736381	1

137 rows × 12 columns

Figure 3-11: Prepared table for final classification

Modeling

For the Final Classification, like for the audio part, we used Binary Classification in the form of TabularModel [14] from PyTorch. In PyTorch, a Tabular Model is typically implemented as Feed-Forward neural network that takes in tabular data as input and outputs a prediction or a classification. The model is designed to take in both continuous and categorical variables as inputs.

The categorical variables are "engagement" and "structured_answers", which are passed through embedding layers to convert them into continuous values that can be fed into the neural network. The continuous variables such as "excited", "eye contact", "smiling", "friendly", "not stressed", "focused" and "not awkward" are normalized to ensure that they are on the same scale. The Tabular Model architecture typically consists of fully connected layers or linear layers, where each layer is connected to the next through a series of weighted connections as shown in Figure 3-12.

The model architecture consists of three linear layers with batch normalization and dropout applied after each layer to prevent overfitting. The final output layer has two units, corresponding to the two possible outcomes (i.e., recommend hiring or not).

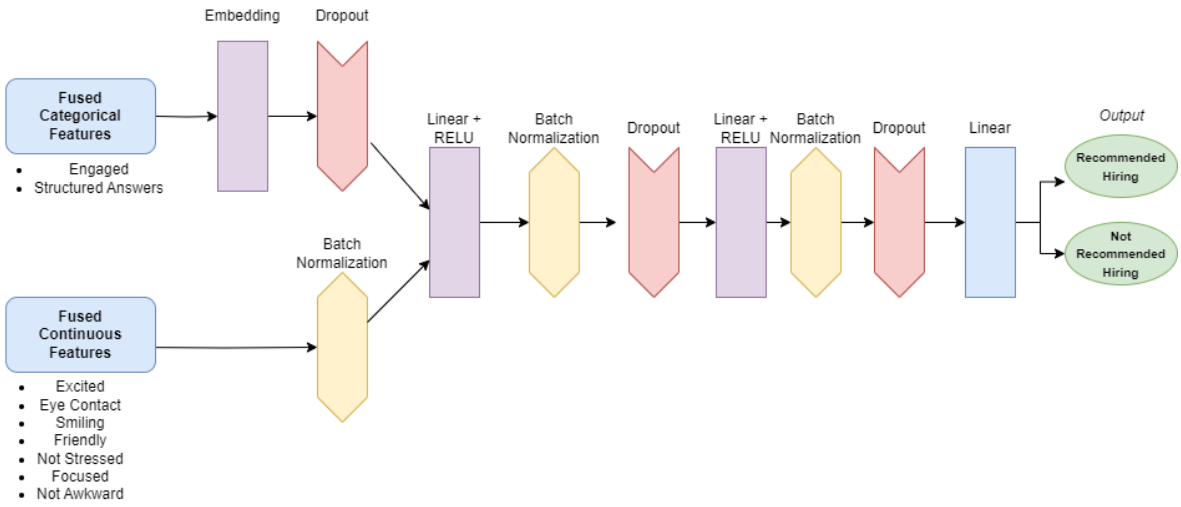

Figure 3-12: FNN for Tabular Data as Final Classification

The model is trained using cross-validation and the Adam optimizer to minimize the binary cross-entropy loss. The embeddings and continuous variables are combined to produce a final prediction for each input.

Overall, this tabular model provides a way to efficiently analyze and predict hiring outcomes based on a combination of continuous and categorical variables.

3.3 Ensembling

Another technique to make final decision was ensembling. Ensembling is a machine learning technique that combines the predictions of multiple models to produce a final prediction. The idea behind ensembling is that by combining the outputs of several models, the ensemble can often achieve better performance than any single model. Architecture of our proposed method does not change. We just alter the final step and have architecture as in Figure 3-13.

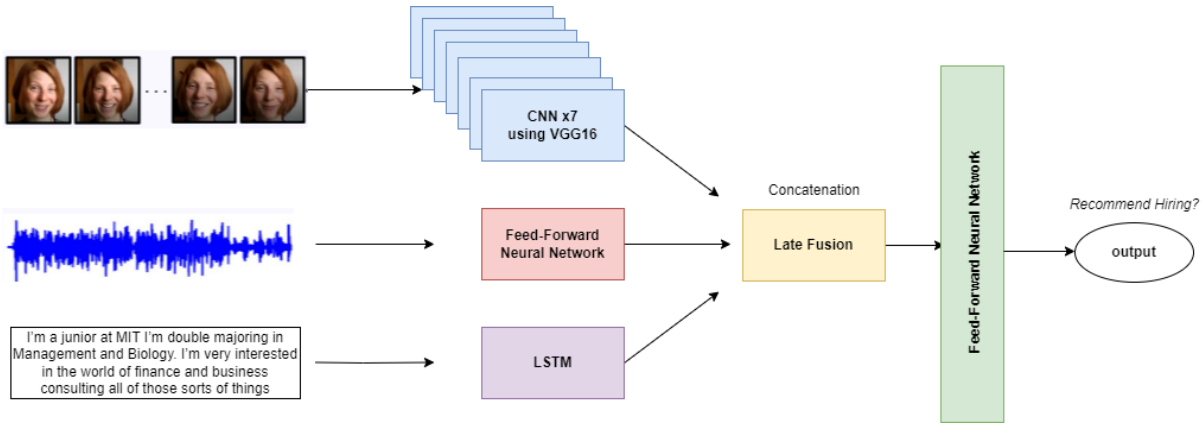

Figure 3-13: Final Decision with Random Forest Regressor

For this purpose, we use Random Forest Regressor. Random Forest Regressor is a regression algorithm that belongs to the family of ensemble learning methods. It is a powerful and popular machine learning technique that combines the results of multiple decision trees to improve the accuracy and robustness of the predictions.

Random Forest is a collection of decision trees where each tree is trained on a random subset of the training data and a random subset of the input features. This randomness helps to reduce the correlation between individual trees and therefore improve the accuracy and robustness of the model.

To make a prediction, each tree in the forest independently produces an output. The final output is then determined by averaging the outputs of all the trees in the forest. This averaging process reduces the variance of the predictions and improves the overall accuracy of the model.

3.4 Evaluation Metrics

Accuracy, Precision, Recall, F1-score, and ROC AUC were the assessment measures we used to assess the performance of our models. The following is how these measurements are expressed:

- Accuracy: The proportion of correctly predicted data to all data is the simplest performance metric to comprehend.

$$\frac{TP + TN}{TP + FN + TN + FP}$$

- Recall: is a metric for how well a model detects True Positives.

$$\frac{TP}{TP + FN}$$

- Precision: it is the proportion of correctly predicted positive observations to all positive observations.

$$\frac{TP}{TP + FP}$$

- F1-score: is the average of Precision and Recall, weighted.

$$\frac{2 * Precision * Recall}{Precision + Recall}$$

where TP, TN, FN, FP are defined respectively as:

(TP) A test outcome that accurately detects the existence of a condition or characteristic.

(TN) A test outcome that accurately demonstrates the absence of a condition or a characteristic.

(FN) A test result that falsely suggests the presence of a certain condition or attribute.

(FP) A test result that falsely suggests the absence of a certain condition or attribute.

Moreover, we evaluate ROC AUC (Receiver Operating Characteristic Area Under the Curve) which is a popular performance metric used in binary classification problems. It measures the area under the ROC curve, which is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.

Chapter 4

Results and Evaluation

Our results are organized in four sections. We first report the prediction accuracies for the trained Deep learning models (CNN, FNN, and LSTM) based on automatically extracted features in section 4.1. In the following section, by comparing the ratings of individual traits to the overall rating, we identify the traits that are most important for success in job interviews. In section 4.3, we compare results with [1] on prediction accuracy. In the last section, we demonstrate the overall system by presenting our front end.

4.1 Prediction Accuracy using Trained Models

To verify our experiment, we used the MIT interview dataset [1], which comprises recordings of 138 simulated job interviews and provide audio and transcripts extracted from these videos. With an allocation of 80% for training, 10% for testing, and 10% for validation, we decided to split the dataset into three separate sets for training, testing, and validation. In each scenario, we train the model on a portion of the dataset, validate it on the validation set, then report on its accuracy on the test set.

Given the feature vectors associated with each interview video, we would like to provide insights into personality traits that are relevant for job interviews and feedback to users about their overall performance in the interview by predicting the likelihood of getting an offer. Overall, we predict 10 labels - 'friendly', 'focused',

	Emotion	Accuracy	Precision	Recall	F1-Score	AUC
1	Eye_contact	0.86	0.80	1.00	0.89	0.89
2	Not_Awkward	0.75	0.68	0.94	0.79	0.85
3	Friendy	0.91	0.87	0.94	0.91	0.90
4	Smiling	0.90	0.92	0.7	0.90	0.93
5	Excited	0.88	0.92	0.81	0.86	0.87
6	Focused	0.91	0.88	0.97	0.92	0.94
7	Not_Stressed	0.90	0.85	0.99	0.92	0.97

Table 4.1: Performance of Emotion recognition from Video Frames.

'not awkward', 'eye contact', 'excited', 'not stressed', 'smiling', 'engaged', 'structured answers' and 'recommend hiring'.

Traits like 'friendly', 'focused', 'not awkward', 'eye contact', 'excited', 'not stressed', and 'smiling' are predicted from visual content. As some of the emotions are very similar to each other, there are separate CNN models trained for each emotion.

In Table 4.1, we have created classification reports containing f-1 score, recall, precision, accuracy, and AUC for each label predicted from visual modality. On average, the visual modality gives 87.3% accurate prediction.

In order to test the video for the aforementioned emotions, we created a function that iterates through each image and loads pre-trained models(saved from the training of each label) for different emotions to predict their emotional content. It tracks predictions for each emotion and by dividing its number by total frames it calculates the percentage of each emotion in total and returns a pandas data frame containing these results.

In order to predict whether a candidate is "engaged" or "not engaged" based on audio recordings of the interview, we trained a Feed-Forward Neural Network using the extracted features of each segmented audio file, including 'pause_number', 'avg_pause_length', 'rhythm_mean', and 'power_mean'. Feed-Forward Neural Network based on the Tabular model was applied. The model has three layers: two linear layers with batch normalization and ReLU activation functions, and one final linear layer with a bias term. The choice of the Feed-Forward Neural Network is suitable for this problem since we have a small number of continuous input features. Our model

Label	Accuracy	Precision	Recall	F1-Score	AUC
Engagement	0.81	0.85	0.90	0.88	0.76

Table 4.2: Audio classification result

achieved an accuracy of 81%, indicating that it can accurately distinguish between engaged and not engaged candidates based on their audio recordings.

For the text, we have used Long Short-Term Memory (LSTM) layers for performing sentiment analysis on candidate answers. The model takes as input a sequence of words and outputs a single value representing the sentiment score of the text. The LSTM model was chosen due to its effectiveness in handling sequential data such as text, and its ability to capture the context and meaning of the words in the sequence. The LSTM layers in the model allow it to handle long-term dependencies in the text, which is important for accurately analyzing the sentiment of longer texts. The model has achieved an accuracy of 83% in this task, which indicates its effectiveness in identifying structured answers.

Label	Accuracy	Precision	Recall	F1-Score	AUC
Structured Answers	0.83	0.93	0.78	0.85	0.90

Table 4.3: Text classification result

The MIT Interview dataset [1] contains evaluations from Amazon Mechanical Turk Workers for each video, which are aggregated to determine the final score for each label and given in CSV file. Since the dataset has different metrics and rates everything from 1-7, we need to divide it by 7 to get the range from 0-1. This will put our predictions to the same representation. The only column that is not going to be altered is "recommend_hiring". Instead, it will use a threshold (> 5 = passed, not passed otherwise) for the final classification.

As mentioned earlier, all of our results from different modalities are fed into one common data frame which will be similar to annotated file with human-rated results. For the final classification after late fusion, Feed-forward Neural Network based

on Tabular Model was trained with target label "Recommend Hiring". The model achieved an accuracy of 92.3% on the test set, indicating that it can accurately predict whether a candidate should be recommended for hiring or not. The five-fold cross-validation results are shown in Table 4.4. The average accuracy across all 5-folds was 91.6%.

Fold	Accuracy
1	0.89
2	0.93
3	0.96
4	0.88
5	0.92
Average	0.916

Table 4.4: 5-Fold Cross Validation for Final Classification

4.1.1 Ensembling with Random Forest

If for the FNN Final prediction we prepared the annotated data putting threshold for the target column 'recommend hiring' in the preprocessing step, for the Random Forest Regressor we take original intensity of recommending column as a continues variable. A random forest regressor model is instantiated with 100 estimators and a random state of 42 using the RandomForestRegressor class from the sklearn.ensemble module. This model loads data from a CSV file and performs 5-fold cross-validation to evaluate its performance. The data is split into 5 sets of training and validation data, with a Random Forest regressor trained on each set of training data and used to make predictions on the corresponding validation data.

The accuracy of each fold is then averaged to give an overall validation accuracy of 93.3% (Table 4-5). This approach provides a more robust evaluation of the model's performance than a simple train-test split and helps to ensure that the model is not overfitting to the training data.

Finally, the model is evaluated using R^2 score metrics. The R^2 score measures the proportion of the variance in the target variable that is predictable from the independent variables. Our regressor gives 0.74 for R^2.

Fold	Accuracy
1	0.92
2	0.925
3	0.936
4	0.941
5	0.942
Average	0.933

Table 4.5: 5-Fold Cross Validation for Ensembling

The output from the regressor is float number which is considered as success rate of the candidate. By this information, we propose to use threshold 60% as the passing the job interview. This will help hiring process by adding more information and comparison between candidates.

In addition to performing 5-fold cross-validation, this model also includes a separate testing set to provide an additional evaluation of its performance. After training the Random Forest regressor on each fold of the training data, the model is evaluated on the testing set to assess its ability to generalize to new, unseen data. On the testing, Random forest Regressor gives 94% accuracy. Following figure illustrates the recommended candidate with his success rate:

```
[0.71127371]
Candidate P52 is recommended to get hired with 71.13%!
```

Figure 4-1: Recommended Candidate with success rate

4.2 Correlation of the Behavioral Traits

The Pearson correlation coefficient is a measure of the linear relationship between two variables. For a pair of variables X and Y, the Pearson correlation coefficient (denoted as r) is calculated as:

$$r = \frac{cov(X, Y)}{std(X) * std(Y)}$$

where cov(X,Y) is the covariance between X and Y, and std(X) and std(Y) are the standard deviations of X and Y, respectively.

4.2.1 Using FNN

We are looking for traits that have a high correlation with ratings. In order to do that, we calculate correlation of each label with the target "recommend hiring" column. In the context of a Pandas DataFrame, DataFrame.corr() computes the pairwise correlation coefficients between all pairs of columns in the DataFrame using the Pearson correlation coefficient.

This information can help job interviewees better understand the most important traits to look for in a job interview. We plot the correlation of labels to the target label "Recommend Hiring" in Figure 4.2,

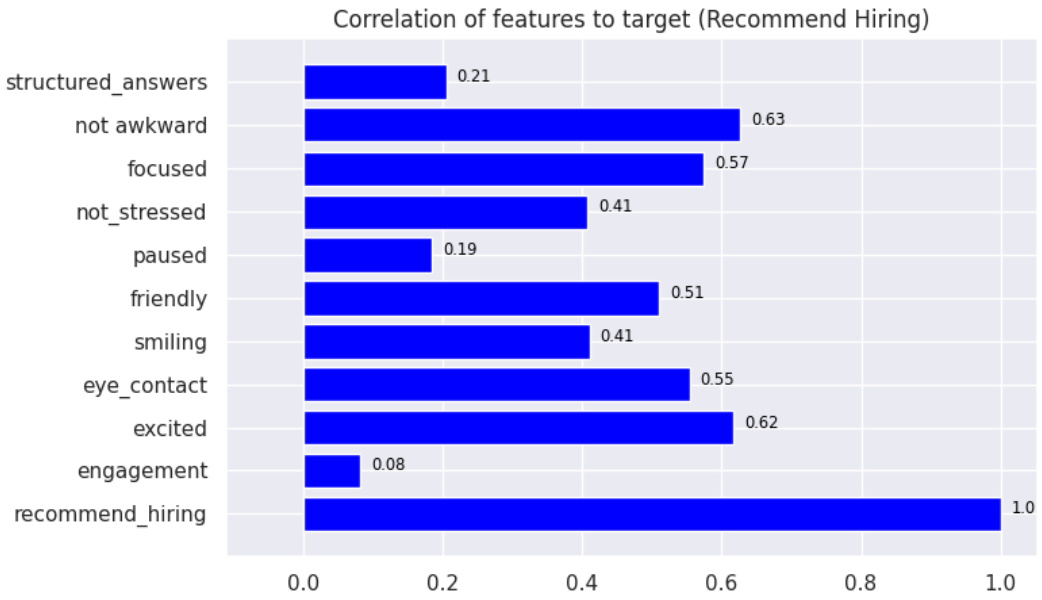

Figure 4-2: Correlation of labels

The plot clearly shows that in an interview, the most important trait is not to appear awkward. The "not awkward" label correlates with the final decision by 63%. Some other top traits include being excited(62%), staying focused(59%), maintaining eye contact(56%), and expressing friendliness(53%). This plot provides us with an insight into the qualities of a successful interview.

4.2.2 Using Random Forest

Random Forest Regressor was also checked for correlation of labels with predicted outcome and showed better results. On average, 20% improvement is observed. The correlation output provides information about how strongly each feature is associated with the predicted outcome (recommend_hiring). A positive correlation means that as the value of the feature increases, the predicted outcome also tends to increase, and a negative correlation means that as the value of the feature increases, the predicted outcome tends to decrease.

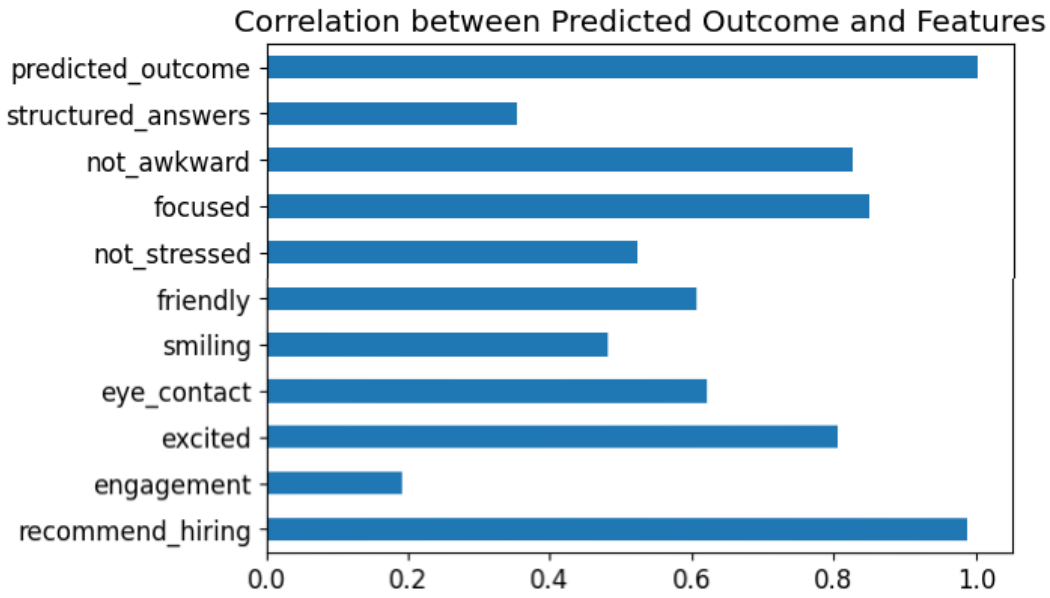

Figure 4-3: Correlation of labels

Looking at the output in Figure 4-3, we can see that the predicted outcome has strong positive correlations with columns such as focused, not_awkward, and excited. This suggests that candidates who were more focused, less awkward, and more excited during the interview tended to receive higher predicted outcomes.

On the other hand, we can see that the predicted outcome has a weak positive correlation with the engagement column and structured_answers column. This suggests that while engagement and structured answers are important factors in an interview, they may not be as strongly correlated with the overall performance and predicted outcome.

4.2.3 Case Study

As previously mentioned, engagement and structured answers are important factors in an interview. Correspondingly, setting a threshold for the engagement and structured_answers columns can be useful in identifying candidates who may not be suitable for the job. If a candidate has a score of 0 in both columns, it could suggest a lack of interest or preparation for the job, and it may be reasonable to exclude them from consideration. In order to realise this, simple conditions were applied after getting

results from individual modalities. For example, let us consider following case:

emotion	engagement	excited	eye_contact	smiling	friendly	paused	not_stressed	focused	speaking_rate	not_awkward	structured_answers
0	0.0	0.7	1.0	0.1	1.0	0.123	0.9	0.1	0.420993	1.0	0.0

Figure 4-4: Case study of Candidate 1

As you can see from the results, all other behavioural traits show high results in average, so by previously mentioned correlations, this candidate had to be hired. But we applied corresponding conditions and now we first look at engagement and structured answers and make following conclusion of excluding from consideration:

Final Prediction

Candidate 1 is not recommended to get hired!

index	engagement	excited	eye_contact	smiling	friendly	paused	not_stressed	focused	speaking_rate	not_awkward	structured_answ
0	0.0	0.7	1.0	0.1	1.0	0.123	0.9	0.1	0.420993	1.0	0.0

Figure 4-5: Results of Candidate 1

However, it is important to mention that if a candidate scores 0 in one of the engagement or structured_answers columns, it's essential to consider the correlations provided before. So, logic for these cases remains the same as it was previously.

Overall, it's important to consider the correlations provided before and evaluate candidates holistically, taking into account various factors related to the job and the organization's needs. While a low score in the engagement or structured_answers columns could be a red flag, it's essential to consider other factors and evaluate the candidate on a holistic basis.

4.3 Compare Results

In [1], they measure prediction accuracy by the area under the curve (AUC) values for the learned models SVM and Lasso. In order to compare results, we also calculated AUC for each of our traits and present results in Table 4.6.

	Trait	Proposed	SVM[1]	Lasso[1]
1	Not Stressed	96.7	60.4	57.2
2	Recommend Hiring	95.2	81.5	79.6
3	Focused	94.5	79.1	67.7
4	Smiling	93.7	84.5	84.5
5	Structured Answers	90.2	81.2	79.9
6	Friendly	89.7	82.4	79.3
7	Eye Contact	89.0	67.6	62.2
8	Excited	87.2	90.4	88.5
9	Not Awkward	82.5	80.8	78.7
10	Engagement	75.9	85.8	85.0

Table 4.6: Comparing Results with [1] based on ROC AUC

The table clearly shows that our model outperformed in AUC for almost all of the traits. For labels Not Stressed, Recommend Hiring, Focused, Smiling and Structured answers, high accuracy was observed exceeding 90%. For some of the labels, there is a huge shift of about 20%, which shows the effectiveness of the proposed model.

4.4 Custom Datasets

4.4.1 Custom Actors Dataset

In order to assess the robustness and efficacy of the Proposed Model of Multimodal Interview Analysis, evaluating it on a custom dataset is a crucial step. Apart from the aforementioned custom dataset, we also tested the proposed approach using 5 videos extracted from various YouTube channels where individuals spoke English to a camera. The purpose was to evaluate the system's capability to correctly identify the final prediction and measure its robustness on a novel and unseen dataset. These results are presented in Table 4.7.

Traits	1	2	3	4	5
Engaged	true	false	false	true	false
Excited	0.7	0.8	0.1	0.8	0.5
Eye_Contact	0.6	0.7	0.3	0.6	0.2
Smiling	0.7	0.5	0.3	0.6	0.2
Friendly	0.4	0.5	0	0.5	0.4
Paused	0.4	0.5	1	0.5	0.6
Not_Stressed	0.5	0.8	0.4	0.3	0.4
Focused	0.6	0.3	0.7	0.7	0.3
Not_Awkward	0.7	0.8	0.2	0.6	0.4
Structured Answers	true	true	true	false	false
Recommend Hiring	Yes	Yes	No	No	No

Table 4.7: Custom Dataset Results

However, even if the data was not labeled, we looked through the videos by ourselves and realized that sometimes the model may show inaccurate results. It can be because the model trained on a specific dataset based on interview content and the custom dataset's videos are just random actors telling different things. But overall, in this testing set system gives 70% accuracy on average.

4.4.2 Custom interview dataset

The process of evaluating the Proposed Model of Multimodal Interview Analysis on a custom dataset is an essential step towards ensuring the system's robustness and effectiveness. In addition to the previously discussed custom dataset, we shot 5 interview videos ourselves, using the same 5 general questions used in the MIT interview dataset. The aim was to test the system's ability to accurately identify the final prediction and evaluate its robustness in a new dataset which is close to the

Label/#	Test1	Test2	Test3	Test4	Test5
Engage-ment	1	1	0	0	1
Excited	0.8	0.4	0.4	0.2	0.3
Eye Contact	0.7	0.8	0.6	0	0.6
Smiling	0.6	0.5	0.4	0.2	0.1
Friendly	0.7	0.5	0.5	0.1	0.2
Not Stressed	0.5	0.4	0.3	0.2	0.3
Focused	0.5	0.6	0.7	0.2	0.4
Not Awkward	1	0.8	0.9	0.4	0.2
Stuctured Answers	1	1	1	0	0
Recommend Hiring	Yes	Yes	Yes	No	No

Table 4.8: Predicted outcomes on Custom Interview Dataset

trained dataset.

Based on self evaluation of the custom dataset, predicted outcomes are showing very close results. Which means that interviews with the same content can be analyzed with our proposed method with about 80% accuracy on average. The comparatively high accuracy of the proposed model when analyzing additional videos with the same content could be attributed to several factors. One possible explanation is that the model has learned to identify patterns in the interview content that are consistently associated with certain outcomes. For example, the model may have learned to associate certain verbal cues, such as hesitations or repetitions, with specific emotions or behaviors.

Additionally, the fact that the additional videos were created using the same questions as the MIT interview dataset may have contributed to the high accuracy of the proposed model. If the questions are highly predictive of the outcomes being analyzed, then the model may be able to accurately predict those outcomes regardless of the specific context or individual being interviewed.

Although the results of testing the Proposed Model of Multimodal Interview Anal-

ysis on the custom dataset were generally promising, it is worth noting that there were some instances where the predicted labels were slightly incorrect. This indicates that the model may not be completely accurate in predicting outcomes in every situation. This limitation could be attributed to the bias in the dataset used for training the system. The model may be limited by potential biases in the training data. The custom dataset may contain biases in terms of race, gender, accent, or other characteristics, and the model may learn to associate certain interview content with specific outcomes based on those biases. This could impact the model's effectiveness and accuracy in real-world applications.

Overall, while the proposed model of multimodal interview analysis shows promising results on the custom dataset, it is important to consider these potential limitations when evaluating its effectiveness and reliability in real-world applications. Further research and testing is needed to address these limitations and ensure that the model is robust and effective across different contexts and populations.

4.5 Demonstration

4.5.1 Running Example

To better understand how the Proposed Model of Multimodal Interview Analysis works in practice, let's consider a running example. In this example, we are analyzing the interview of a candidate with candidate id "P52".

The first step in the analysis is to use the CNN component of the model to detect the candidate's facial emotions during the interview. The CNN analyzes the video footage of the interview and identifies facial expressions such as excited, eye contact, smiling, friendly, not stressed, focused and not awkward. In order to test the video for the aforementioned emotions, we created a function that iterates through each image and loads pre-trained models(saved from the training of each label) for different emotions to predict their emotional content. It tracks predictions for each emotion and by dividing its number by total frames it calculates the percentage of each

emotion in total and returns a pandas data frame containing these results. Results from facial modality illustrated in Figure 4-6.

	emotion	number
0	excited	0.607143
1	smiling	0.410714
2	not_stressed	0.446429
3	eye_contact	0.232143
4	not_awkward	0.928571
5	friendly	0.250000
6	focused	0.017857

Figure 4-6: Facial modality results

Next, we use the FNN component to analyze the audio extracted rom the interview. The FNN processes extracted features such as pause number, average pause length, power mean and rhythm mean of the voice to get the prediction on whether the candidate expressed an engaged tone during the interview. The candidate "P52" showed engaged tone:

```
audio_prediction = get_preds(preds)
audio_prediction
```

```
1
```

Figure 4-7: Audio modality result

Finally, we use the LSTM component to analyze the candidate's text-based responses to interview questions. Transcript from the interview was analyzed to make a prediction on whether the candidate's answers were well-structured and organized, indicating strong communication skills and the ability to convey information effectively. The candidate "P52" had structured answers:

```
prediction = predict(net, test, seq_length)

Prediction value: 1
structured answers
```

Figure 4-8: Text modality result

After analyzing each modality, we concatenate the results into a single dataframe. This concatenated dataframe contains the analysis of the candidate's facial expressions, tone of voice, and text-based responses during the interview. Figure 4-9 shows concatenated results.

```
all_results
```

	emotion	engagement	excited	eye_contact	smiling	friendly	paused	not_stressed	focused	speaking_rate	not_awkward	structured_answers
0		True	0.607143	0.232143	0.410714	0.25	0.285714	0.446429	0.017857	0.573237	0.928571	True

Figure 4-9: Fusion result

We then feed this dataframe into the final fusion and classification component of the model, which uses a Feedforward Neural Network (FNN) to classify the candidate as either recommended or not recommended for the position. The output of the FNN represents the model's recommendation for hiring the candidate based on the combined analysis of their facial expressions, tone of voice, and text-based responses.

```
Candidate P52 is recommended to get hired!
```

Figure 4-10: Final Classification based on FNN

In addition to the FNN, an ensembling technique using a Random Forest Regressor was also incorporated in the final classification component of the Proposed Model of Multimodal Interview Analysis. This ensemble model takes into account the concatenated data from each modality and generates a recommendation for hiring with intensity.

```
[0.71127371]
Candidate P52 is recommended to get hired with 71.13%!
```

Figure 4-11: Final Classification based on Random Forest Regressor

4.5.2 Frontend

For convenience and for visualization of the results, we created a custom interactive web app using Python library Panel [22].

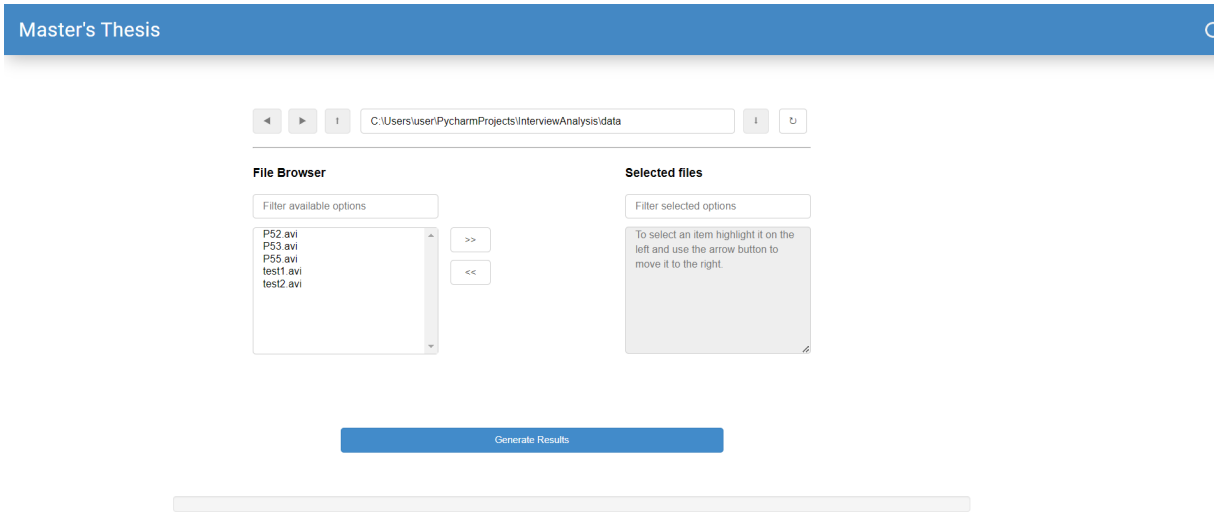

Figure 4-12: Front-end

We created a button to upload a video (Figure 4-13). Unlike previous methods, the system will automatically convert video to audio using the Python library MoviePy [23]. This library is a versatile tool for working with videos, allowing users to perform a wide range of editing tasks, including the extraction of audio. By using MoviePy to convert the video file to an audio file, the transcription process is streamlined and made much more straightforward.

The next step in the process is to convert the audio into text using the SpeechRecognition library provided by Google Cloud [24]. This library offers a user-friendly API for converting speech to text and allows for the selection of various speech recognition

engines. In this case, Google Cloud's speech-to-text API was selected for its accuracy and efficiency.

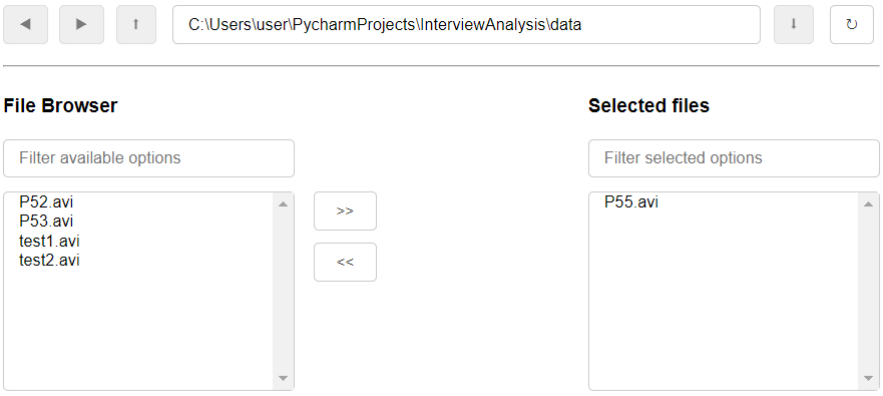

Figure 4-13: Uploading the video

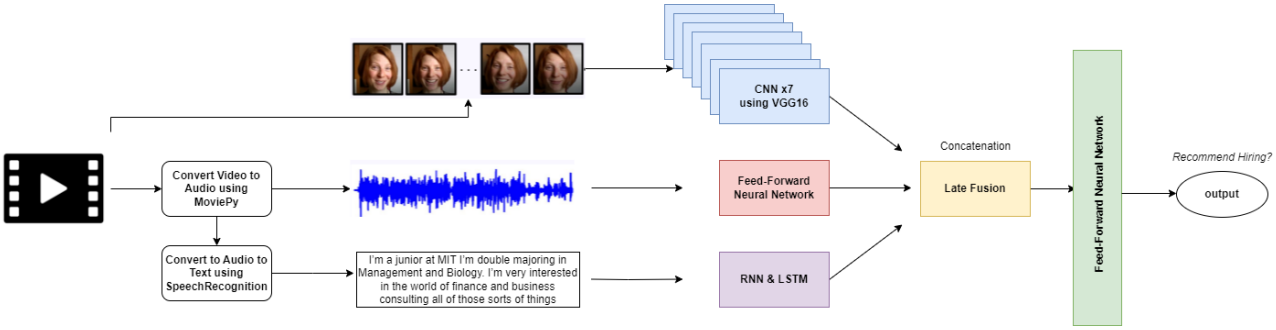

Figure 4-14: Process of getting results

After all the data is prepared, the data runs through the pipeline. All the labels are predicted and written down in one common data file. As a result, the final classification is applied to the resultant data frame and suggests whether the candidate is "Recommended Hiring" or "Not Recommended Hiring" as outlined in Figure 4-14.

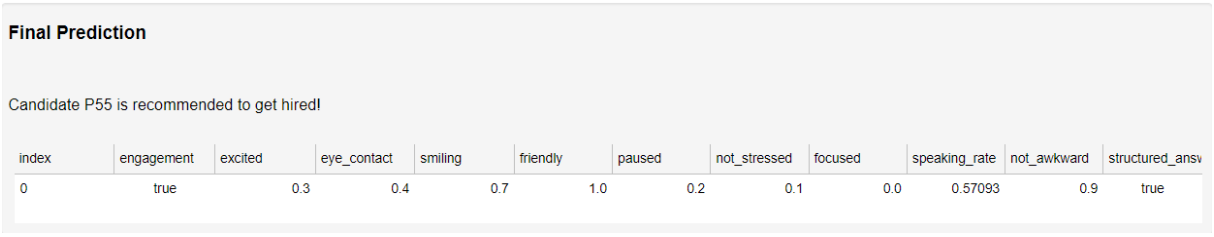

Figure 4-15: Result

Chapter 5

Conclusion

The proposed multimodal interview analysis system has the potential to provide invaluable insights for hiring recommendations by utilizing deep learning techniques. The system utilizes CNNs for visual data, LSTM for text data, and FNN for audio data to capture and process information across various modalities, resulting in accurate predictions with an average accuracy of 87.3%, 83%, and 81%, respectively.

A fusion approach is used to merge information from multiple modalities into one data frame and then use tabular models for prediction of the overall performance of the candidate, resulting in highly successful hiring recommendations with an accuracy of 92.3%, which is higher than any individual modality alone. Ensemble technique is applied using Random Forest Regressor and gives 94% accuracy and demonstrates the intensity of being recommended for hiring.

This approach produced high accuracies for not stressed, recommended hiring, focused, and smiling candidates, with AUCs of over 90% for overall interview analysis. These results exceed previous work, which saw AUCs of approximately 80%, and illustrate that all models proposed in this study reach state-of-the-art performance scores.

Through analyzing the correlation of predicted behavioral traits, the framework recommends certain traits that candidates should display to succeed in job interviews, such as not showing awkwardness, being excited about responses, staying focused, maintaining eye contact, and expressing friendliness. The study also created a cen-

tralized custom interactive UI for showing clear results. Overall, the effectiveness of the proposed system in capturing and analyzing information across various modalities could lead to more informed hiring decisions and has the potential to be applied to other contexts where the analysis of human emotional behavior is relevant.

Moreover, the proposed model of multimodal interview analysis showed promising results when evaluated on a custom interview dataset. The model was able to accurately predict outcomes in interviews with the same content, achieving an average accuracy of about 80%. This high accuracy can be attributed to various factors such as the model's ability to identify consistent patterns in the interview content and the predictive power of the interview questions. However, it's important to acknowledge that there were instances where the predicted labels were slightly incorrect, indicating that further improvements may be necessary to increase the model's reliability and effectiveness. Nevertheless, these findings provide a good foundation for future research and development of the proposed model.

Bibliography

[1] Naim, I., Tanveer, M. I., Gildea, D., & Hoque, M. E. (2018). Automated Analysis and Prediction of Job Interview Performance. IEEE Transactions on Affective Computing, 9(2), 191–204. doi:10.1109/taffc.2016.2614299

[2] Agrawal, A., George, R. A., & Ravi, S. S. (2020). Leveraging multimodal behavioral analytics for automated job interview performance assessment and feedback. arXiv preprint arXiv:2006.07909.

[3] Mishra, R., Barnwal, S. K., Malviya, S., Mishra, P., & Tiwary, U. S. (2020). Prosodic feature selection of personality traits for job interview performance. In Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA 2018) held in Vellore, India, December 6-8, 2018, Volume 1 (pp. 673-682). Springer International Publishing.

[4] Middya, A. I., Nag, B., & Roy, S. (2022). Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities. Knowledge-Based Systems, 244, 108580.

[5] Dai, W., Liu, Z., Yu, T., & Fung, P. (2020). Modality-transferable emotion embeddings for low-resource multimodal emotion recognition. arXiv preprint arXiv:2009.09629.

[6] Jiang, D., Liu, H., Wei, R., & Tu, G. (2023). CSAT-FTCN: A Fuzzy-Oriented Model with Contextual Self-attention Network for Multimodal Emotion Recognition. Cognitive Computation, 1-10.

[7] Kumar, P., Malik, S., & Raman, B. (2022). Interpretable Multimodal Emotion Recognition using Hybrid Fusion of Speech and Image Data. arXiv preprint arXiv:2208.11868.

[8] Abdu, S. A., Yousef, A. H., & Salem, A. (2021). Multimodal video sentiment analysis using deep learning approaches, a survey. Information Fusion, 76, 204-226.

[9] Jackson, P., & Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. University of Surrey: Guildford, UK.

[10] Livingstone SR, Russo FA (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): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[11] Gaw, N., Yousefi, S., & Gahrooei, M. R. (2022). Multimodal data fusion for systems improvement: A review. IISE Transactions, 54(11), 1098-1116.

[12] Ding, N., Tian, S. W., & Yu, L. (2022). A multimodal fusion method for sarcasm detection based on late fusion. Multimedia Tools and Applications, 81(6), 8597-8616.

[13] Gadzicki, K., Khamsehashari, R., & Zetzsche, C. (2020, July). Early vs late fusion in multimodal convolutional neural networks. In 2020 IEEE 23rd international conference on information fusion (FUSION) (pp. 1-6). IEEE.

[14] Ortega, J. D., Senoussaoui, M., Granger, E., Pedersoli, M., Cardinal, P., & Koerich, A. L. (2019). Multimodal fusion with deep neural networks for audio-video emotion recognition. arXiv preprint arXiv:1907.03196.

[15] Li, Y., Wan, J., Miao, Q., Escalera, S., Fang, H., Chen, H., ... & Guo, G. (2020). Cr-net: A deep classification-regression network for multimodal apparent personality analysis. International Journal of Computer Vision, 128, 2763-2780.

[16] Chopra, S., & Urolagin, S. (2020, November). Interview Data Analysis using Machine Learning Techniques to Predict Personality Traits. In 2020 Seventh International Conference on Information Technology Trends (ITT) (pp. 48-53). IEEE.

[17] Khan, M., Chakraborty, S., Astya, R., & Khepra, S. (2019, October). Face detection and recognition using OpenCV. In 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS) (pp. 116-119). IEEE.

[18] Kusuma, G. P., Jonathan, J., & Lim, A. P. (2020). Emotion recognition on fer-2013 face images using fine-tuned vgg-16. Advances in Science, Technology and Engineering Systems Journal, 5(6), 315-322.

[19] Joseph, M. (2021). PyTorch Tabular: A Framework for Deep Learning with Tabular Data. arXiv preprint arXiv:2106.12613

[20] Rehman, A. U., Malik, A. K., Raza, B., & Ali, W. (2019). A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. Multimedia Tools and Applications, 78, 26597-26613.

[21] Xu, G., Li, W., & Liu, J. (2020). A social emotion classification approach using multi-model fusion. Future Generation Computer Systems, 102, 347-356.

[22] Chapagain, A. (2019). Hands-On Web Scraping with Python: Perform advanced scraping operations using various Python libraries and tools such as Selenium, Regex, and others. Packt Publishing Ltd.

[23] Porwal, K., Srivastava, H., Gupta, R., Pratap Mall, S., & Gupta, N. (2022). Video Transcription and Summarization using NLP. Available at SSRN 4157647.

[24] Tseng, J. L. (2021). Intelligent augmented reality system based on speech recognition. International Journal of Circuits, Systems and Signal Processing, 15, 178-186.

[25] Jim Schwoebel (2019). Pauses [Source code]. https://github.com/jim-schwoebel/pauses.

[26] Mujtaba, D. F., Qadir, J., & Lee, C. M. (2018). Multi-task deep neural networks for multimodal personality trait prediction. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (pp. 1733-1736).

[27] Kaya, H., & Salah, A. A. (2019). Multimodal personality trait analysis for explainable modeling of job interview decisions. IEEE Transactions on Affective Computing, 12(1), 98-112.