NAZARBAYEV UNIVERSITY

Multimodal Performance Analysis during job interviews

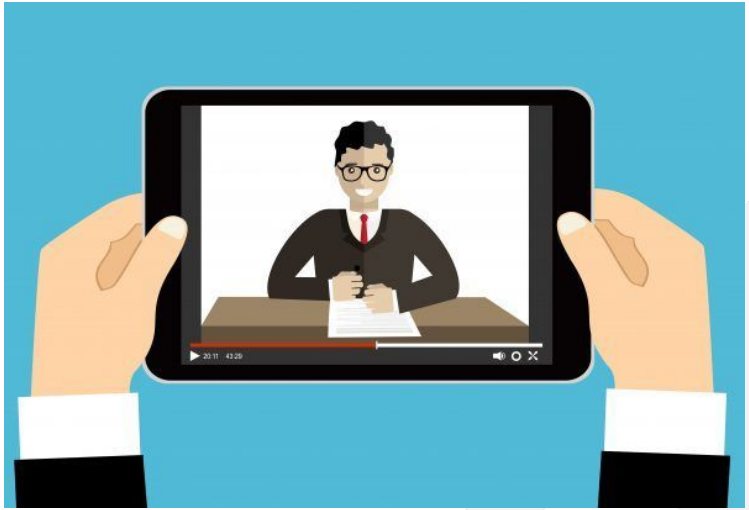
By Amina Aman
Thesis Supervisor: Adnan Yazici, Department Chair of Computer Science at Nazarbayev University

NAZARBAYEV UNIVERSITY

Introduction

- Emotion recognition based on multimodal data has become an important research topic with a wide range of applications, including online interviews.

- To gain a deeper understanding of the interviewee's responses, it is necessary to analyze the interview process multimodally.



NAZARBAYEV UNIVERSITY

Related Works

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

NAZARBAYEV UNIVERSITY

Related Works

Reference	Year	Dataset	Model	Recognized labels	Accuracy
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 Stressed Not Awkward	AUC avg 0.80
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
Chopra et al.	2020	MIT interview dataset	SVR, KNN, Decision Tree	Friendly, Engaged, Excited, Speaking Rate, Calm	AUC avg 0.72

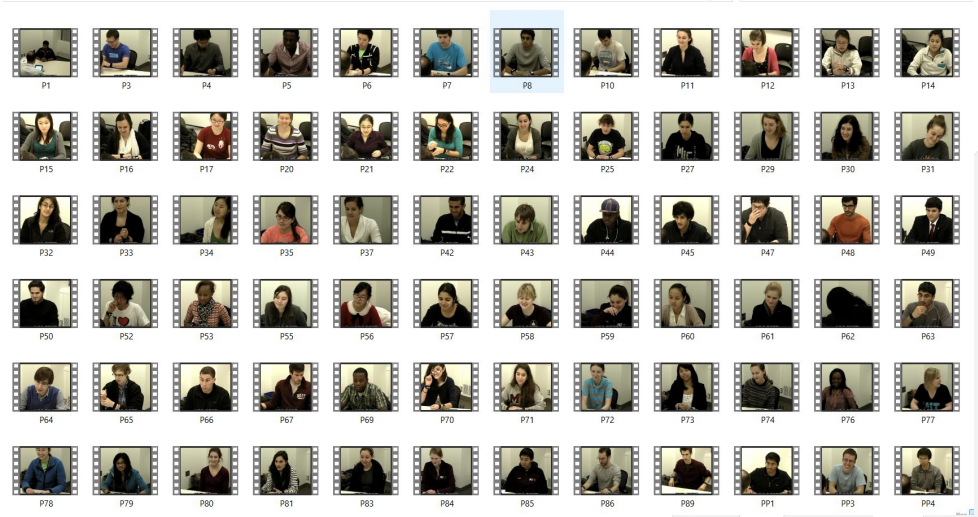
NAZARBAYEV UNIVERSITY

MIT Interview Dataset

- The MIT Interview Dataset contains the audio-visual recordings of 138 mock job interviews, conducted by professional career counselors with 69 undergraduate MIT students.

- Both the video and audio recordings of each interview, text transcripts, and additional annotations by Amazon Mechanical Turk workers are published.

- There are 138 interview videos totaling around 10.5 hours in length, or 4.7 minutes for each interview on average.



NAZARBAYEV UNIVERSITY

Procedure

- During each interview session, the counselor asked interviewees five different questions, which were recommended by the MIT Career Services. These five questions were presented in the following order by the counselors to the participants:

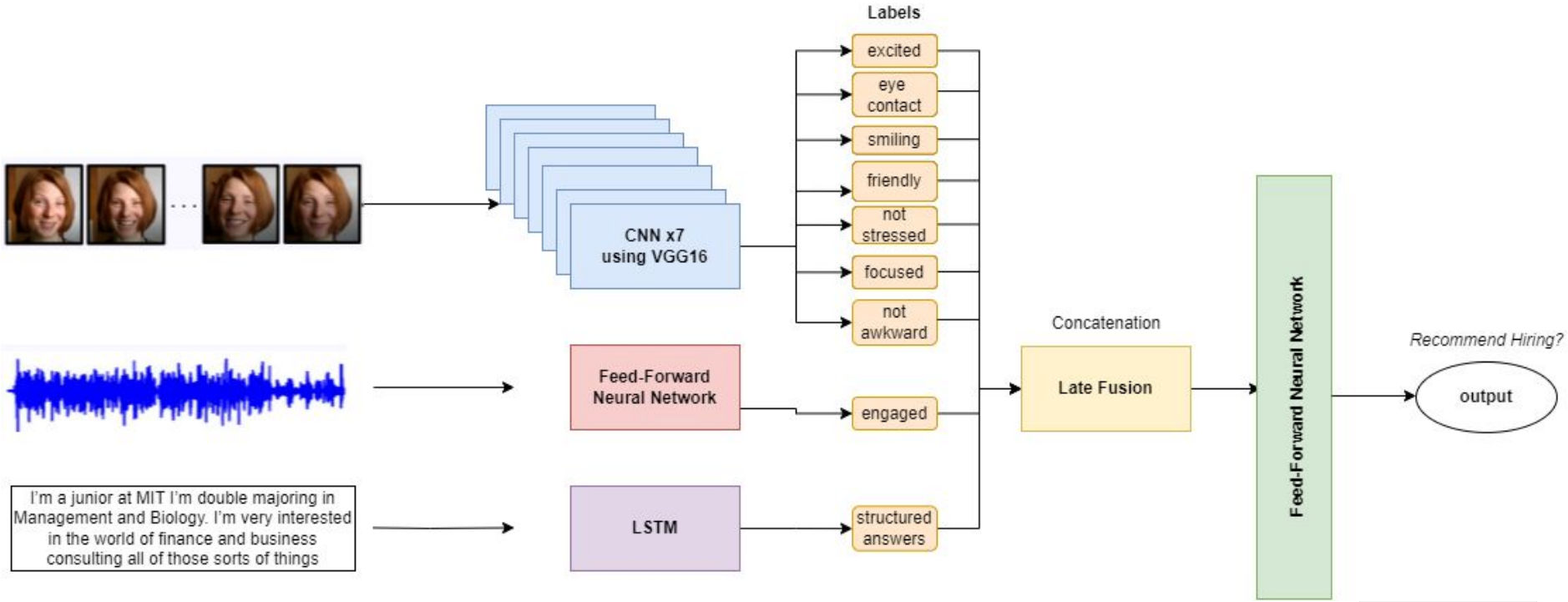
Q1	So please tell me about yourself.
Q2	Tell me about a time when you demonstrated leadership
Q3	Tell me about a time when you were working with a team and faced a challenge.
Q4	What is one of your weaknesses and how do you plan to overcome it?
Q5	Now, why do you think we should hire you?

Table 3.1: List of Interview questions

NAZARBAYEV UNIVERSITY

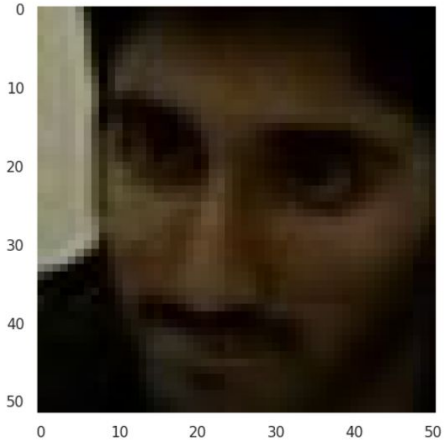
Methodology

- CNN models to classify emotions
- Feed-Forward Neural Network for audio features
- Sentiment Analysis model using LSTM for text
- Late fusion - putting all together in a dataframe
- Feed-Forward Neural Network for Final Classification

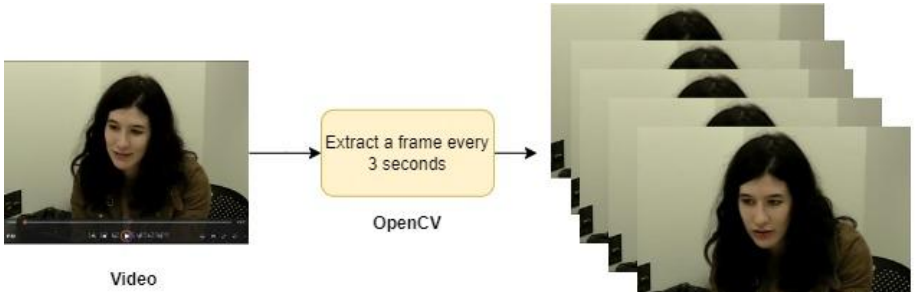


NAZARBAYEV UNIVERSITY

Visual Modality

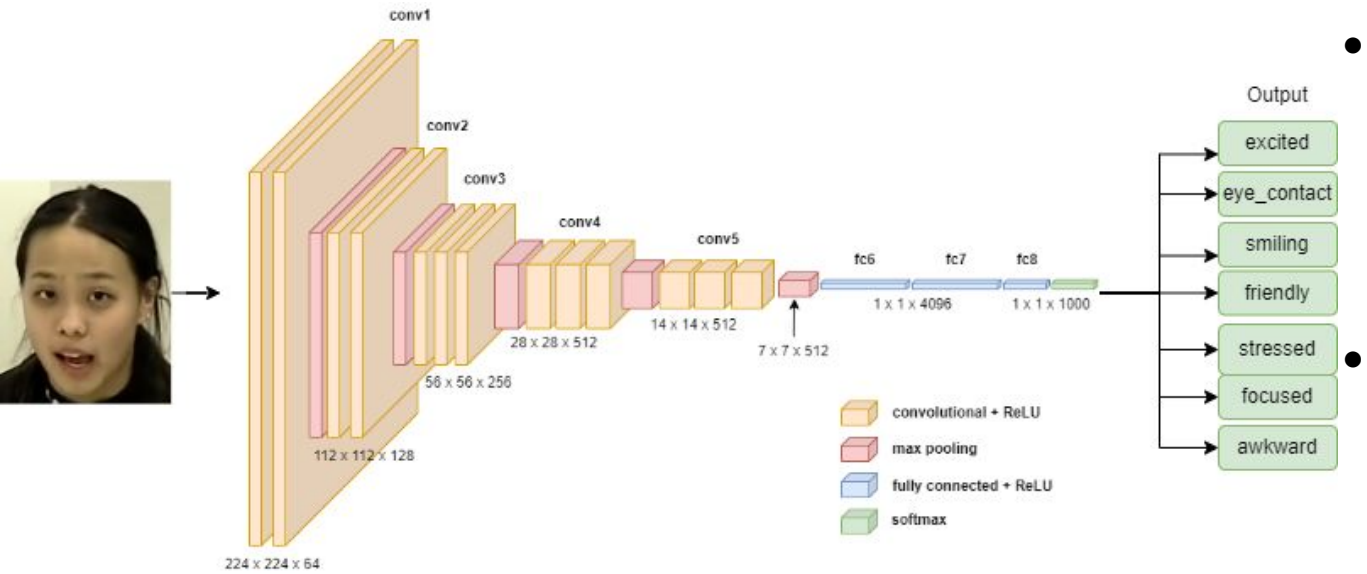


Data preprocessing:



NAZARBAYEV UNIVERSITY

Visual Modality



- For the face part, we designated emotion labels such as ’friendly’, ’focused’, ’awkward’, ’eye contact’, ’excited’, ’stressed’, ’smiling’, which are easy to determine from visual content.
- For that we use transfer learning using the Visual Geometry Group-16 (VGG-16) classification model.

NAZARBAYEV UNIVERSITY

Audio Modality

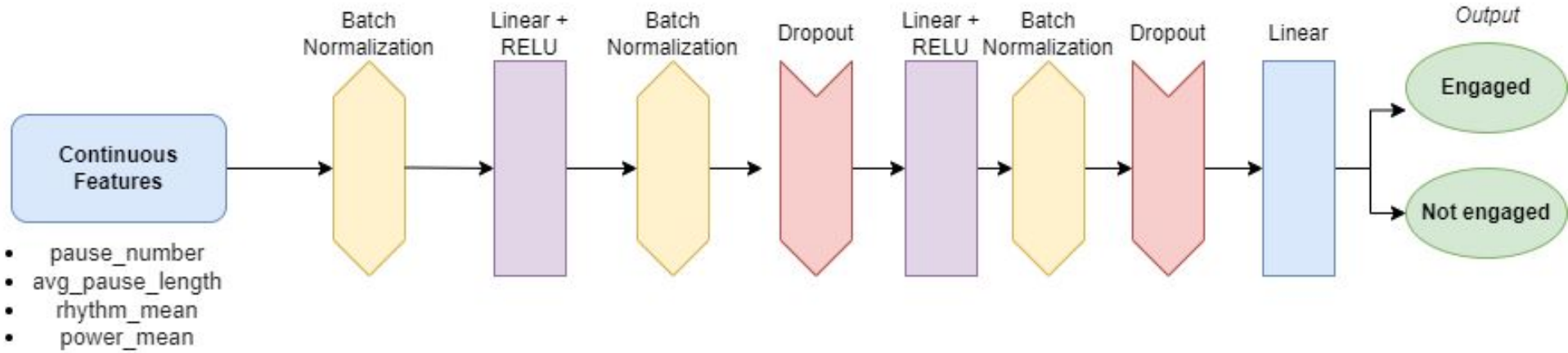
Data preprocessing:

- In MIT Interview Dataset, interviews are in style of dialogue between interviewer and interviewee. Interviewees' responses were extracted and audio were cut into small excerpts according to responses of the candidate.
- To extract features from audio files we used **Librosa library**.
- Mainly "*pause_number*", *"avg_pause_length", "rhythm_mean"* and *"power_mean"* were extracted and written down to csv file.

	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

NAZARBAYEV UNIVERSITY

Audio Modality



- Extracted feature are written down in the tabular format and the target column to predict is taken as the candidate is "engaged" or "not engaged".
- FFN based on TabularModel from PyTorch was chosen to handle dataframe tabular data

NAZARBAYEV UNIVERSITY

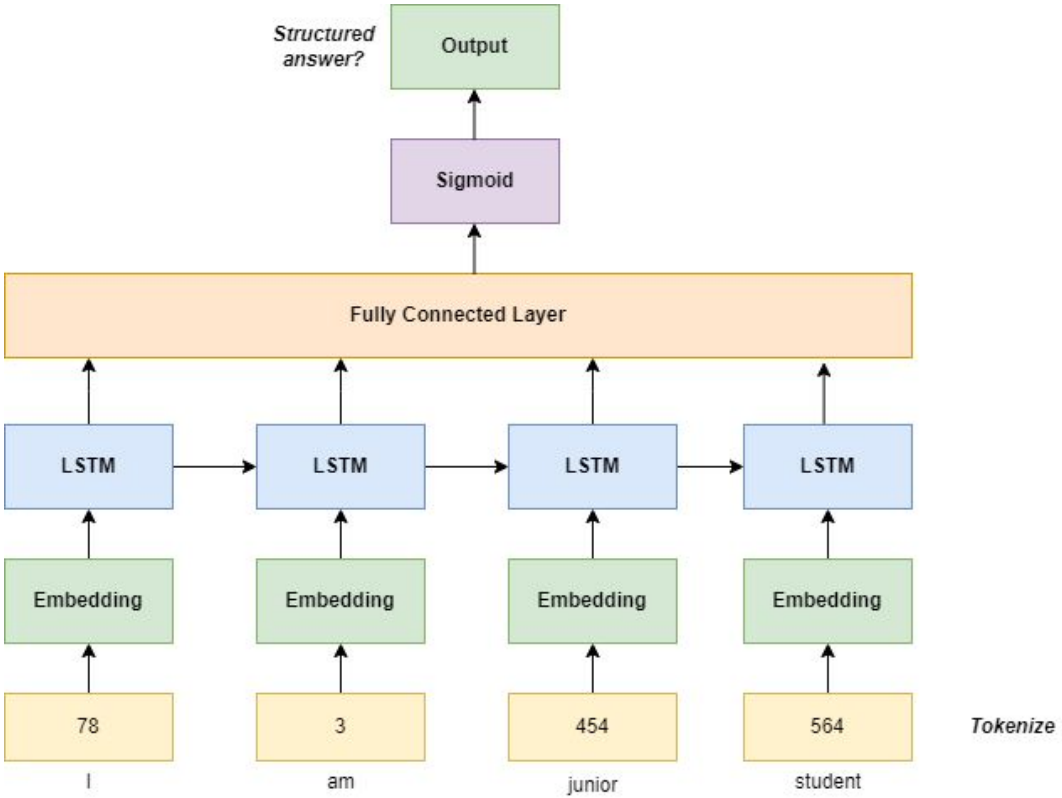
Lexical modality

Data preparation:

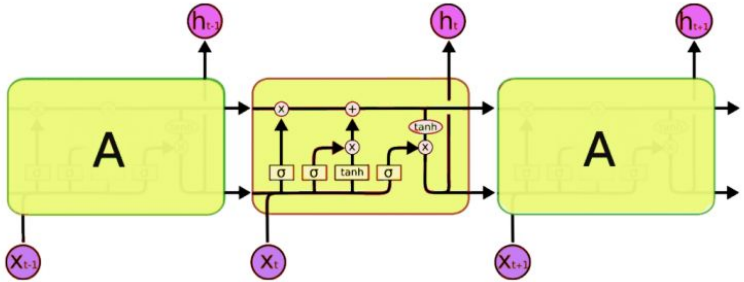
- The transcript is provided in annotated form with the beginning and finish of each interviewee response to help distinguish between the interviewer's and interviewee's speech.
- Accordingly, the interviewers' part was removed from text in order to analyse the candidates' responses.
- We divide full text into sentences, remove punctuation, lower the words and tokenize each sentence.

NAZARBAYEV UNIVERSITY

Lexical modality



- LSTM was used to develop a sentiment analysis model for text classification.
- LSTM Architecture:



NAZARBAYEV UNIVERSITY

Late Fusion and Final Classification

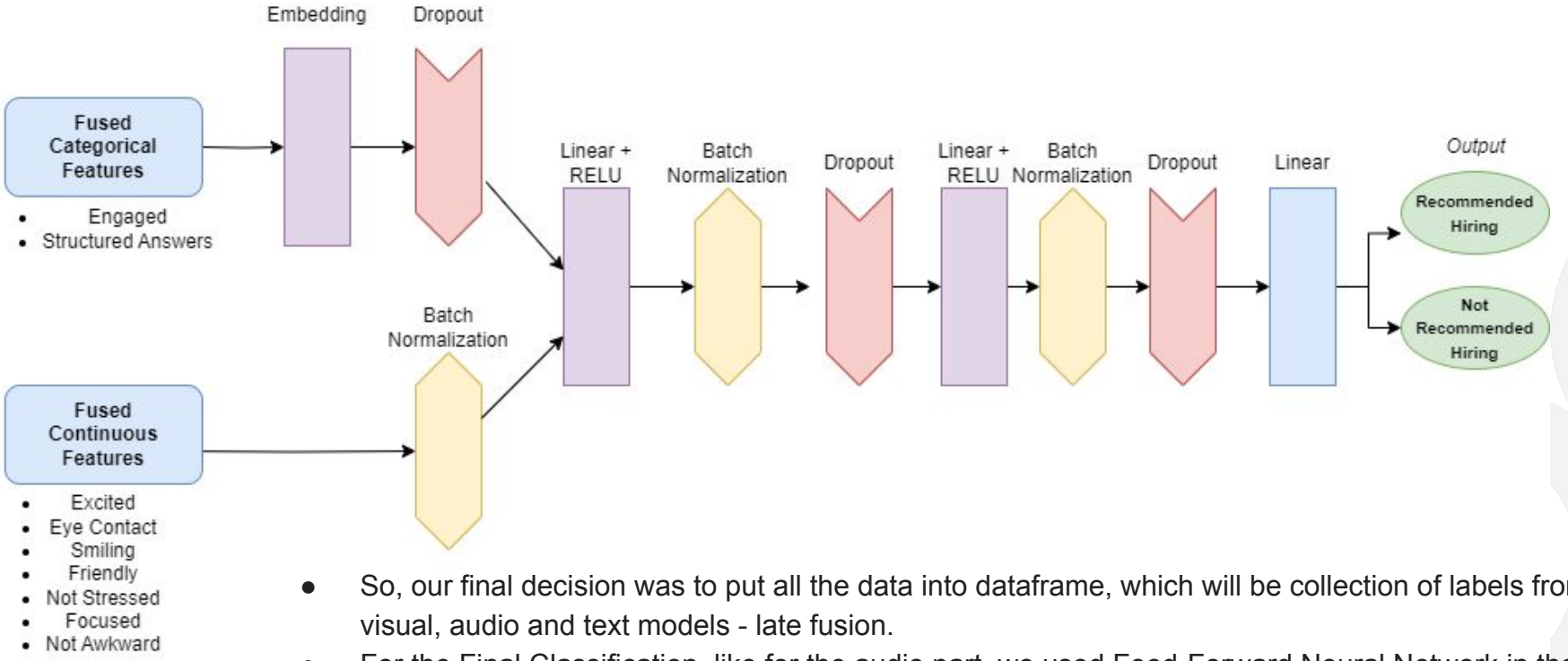
- The MIT Interview dataset [1] contains evaluations from Amazon Mechanical Turk Workers for each video, which are aggregated to determine the final score for each label and given in CSV file. Since the dataset has different metrics and rates everything from 1-7, we need to divide it by 7 to get the range from 0-1. This will put our predictions to the same representation. 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

NAZARBAYEV UNIVERSITY

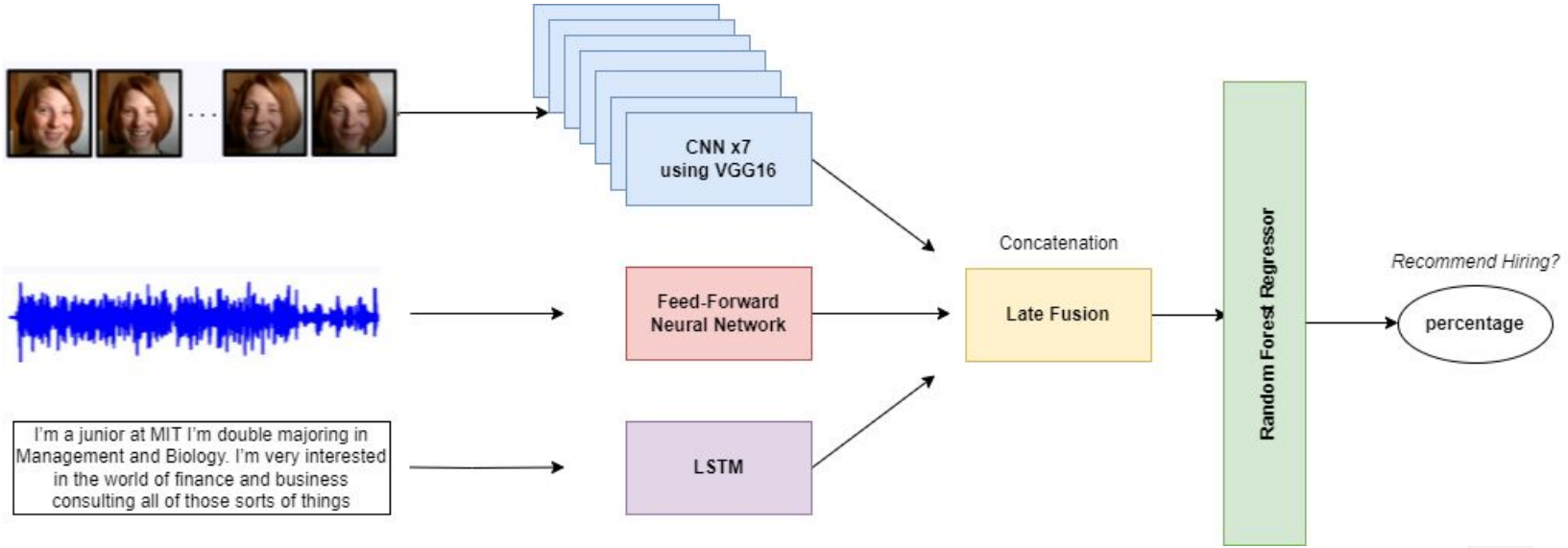
Late Fusion and Final Classification



- So, our final decision was to put all the data into dataframe, which will be collection of labels from visual, audio and text models - late fusion.
- For the Final Classification, like for the audio part, we used Feed-Forward Neural Network in the form of TabularModel from pyTorch.

NAZARBAYEV UNIVERSITY

Ensembling with Random Forest Regressor



- 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.

NAZARBAYEV UNIVERSITY

Results

- Division of the dataset:
  80% for training, 10% for testing, and 10% for validation
- 10 labels - ’friendly’, ’focused’, ’not awkward’, ’eye contact’, ’excited’, ’not stressed’, ’smiling’, ’engaged’, ’structured answers’ and ’recommend hiring’.

	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.

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

Table 4.2: Audio classification result

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

NAZARBAYEV UNIVERSITY

K-Fold Cross Validation For Final Classification

k=5

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

The five-fold cross-validation results are shown in Table. The average accuracy across all 5-folds was 91.6%.

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.

NAZARBAYEV UNIVERSITY

Ensembling Results

- 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.
- This model loads data from a CSV file and performs 5-fold cross-validation to evaluate its performance.

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

- 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%!
```

NAZARBAYEV UNIVERSITY

Results

- For test case let's take video **p52**
- From face we get:

	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

- From audio:

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

```
1
```

- From text:

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

```
Prediction value: 1
structured answers
```

- After late fusion:

```
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

- Final prediction with FNN:

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

- Final prediction with Random Forest:

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

NAZARBAYEV UNIVERSITY

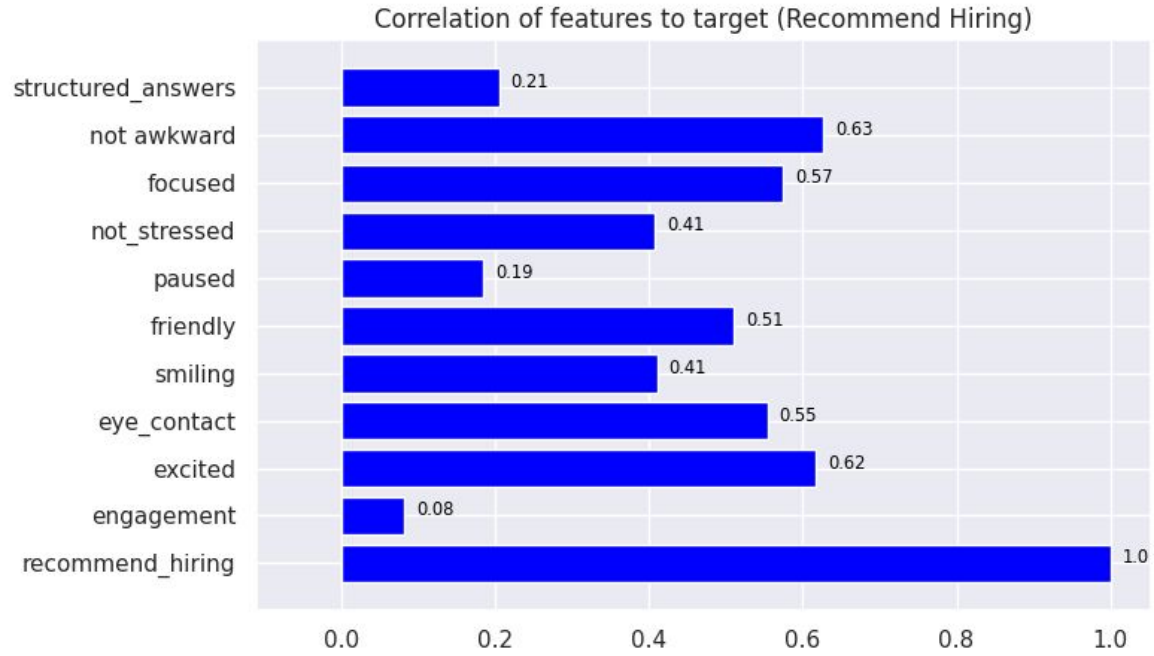
Compare Results

	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

ROC AUC (Receiver Operating Characteristic Area Under the Curve) 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.

NAZARBAYEV UNIVERSITY

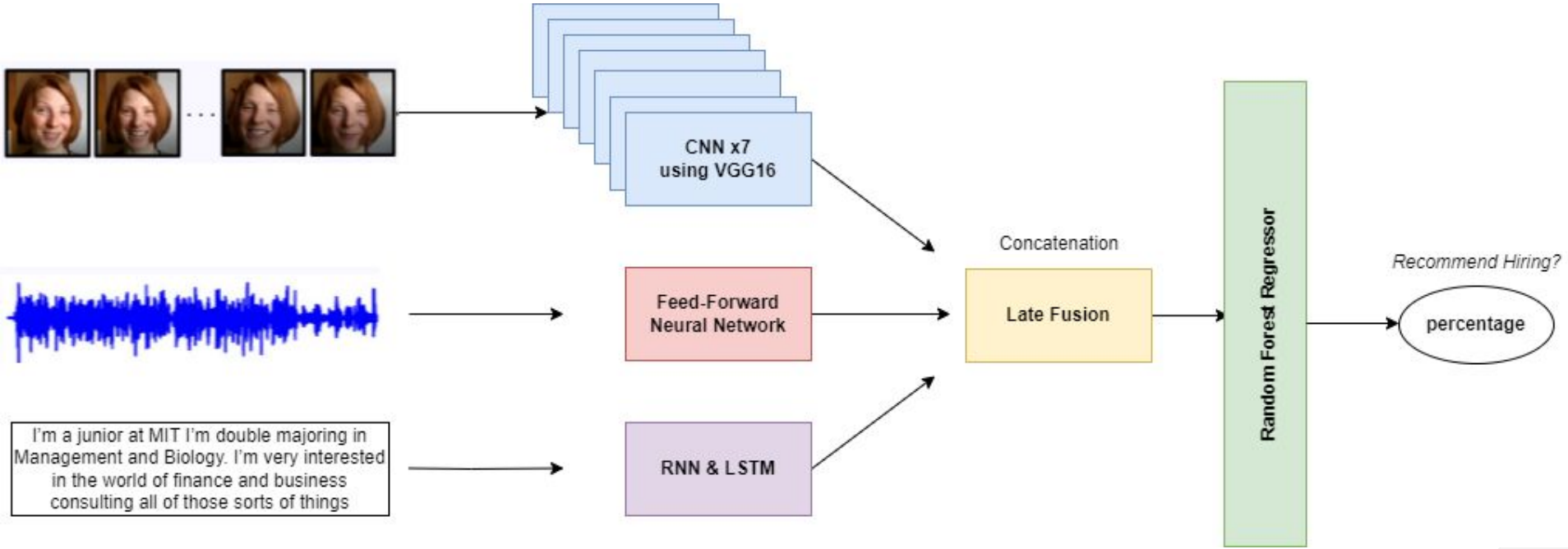
Correlation of the Behavioral Traits



➔ 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.

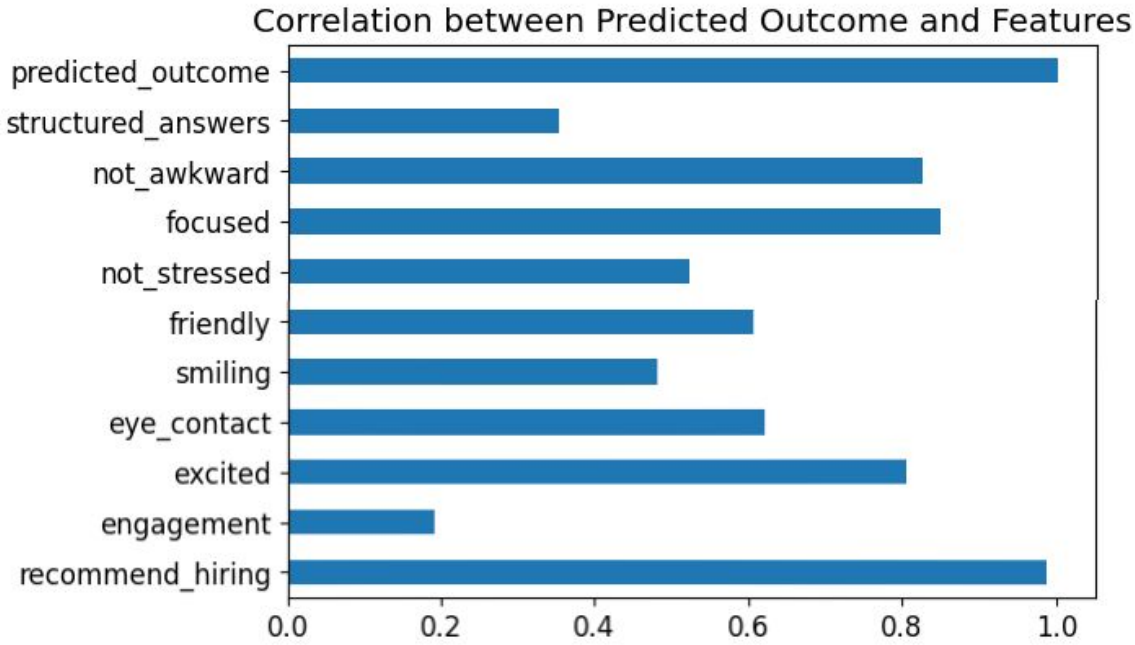
NAZARBAYEV UNIVERSITY

Changing last layer to RandomForestRegressor



NAZARBAYEV UNIVERSITY

Changing last layer to RandomForestRegressor




```
print(outcome_corr)

predicted_outcome      1.000000
recommend_hiring       0.987768
focused                0.847334
not_awkward            0.824525
excited                0.805382
eye_contact            0.622188
friendly               0.605403
not_stressed           0.521779
smiling                0.480498
structured_answers     0.351458
engagement             0.192085
```

NAZARBAYEV UNIVERSITY

Case Study

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

- ➔ 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.*
- ➔ 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.
- ➔ Example:

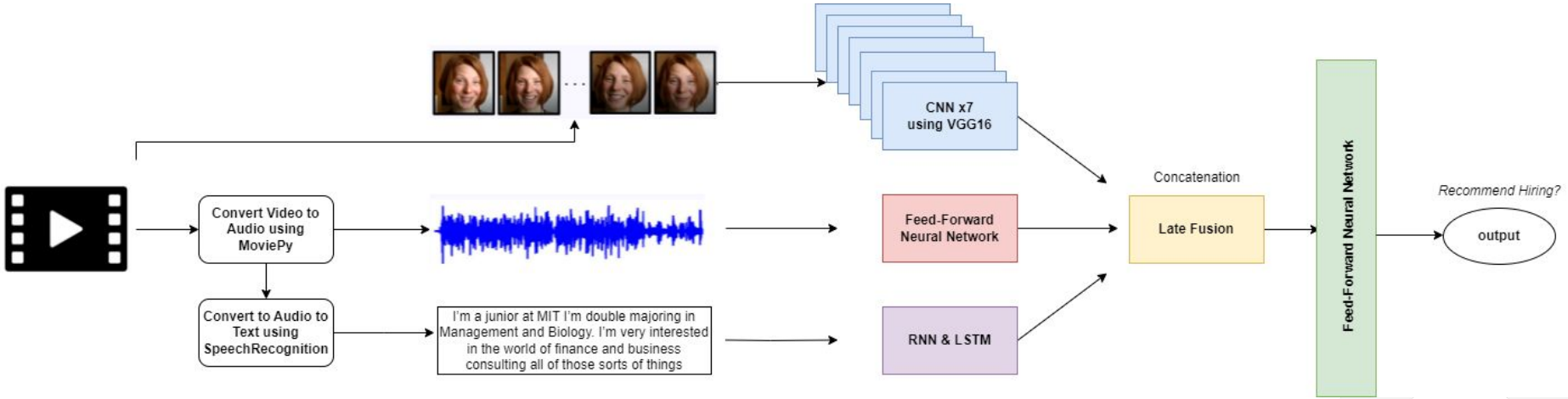
Final Prediction

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

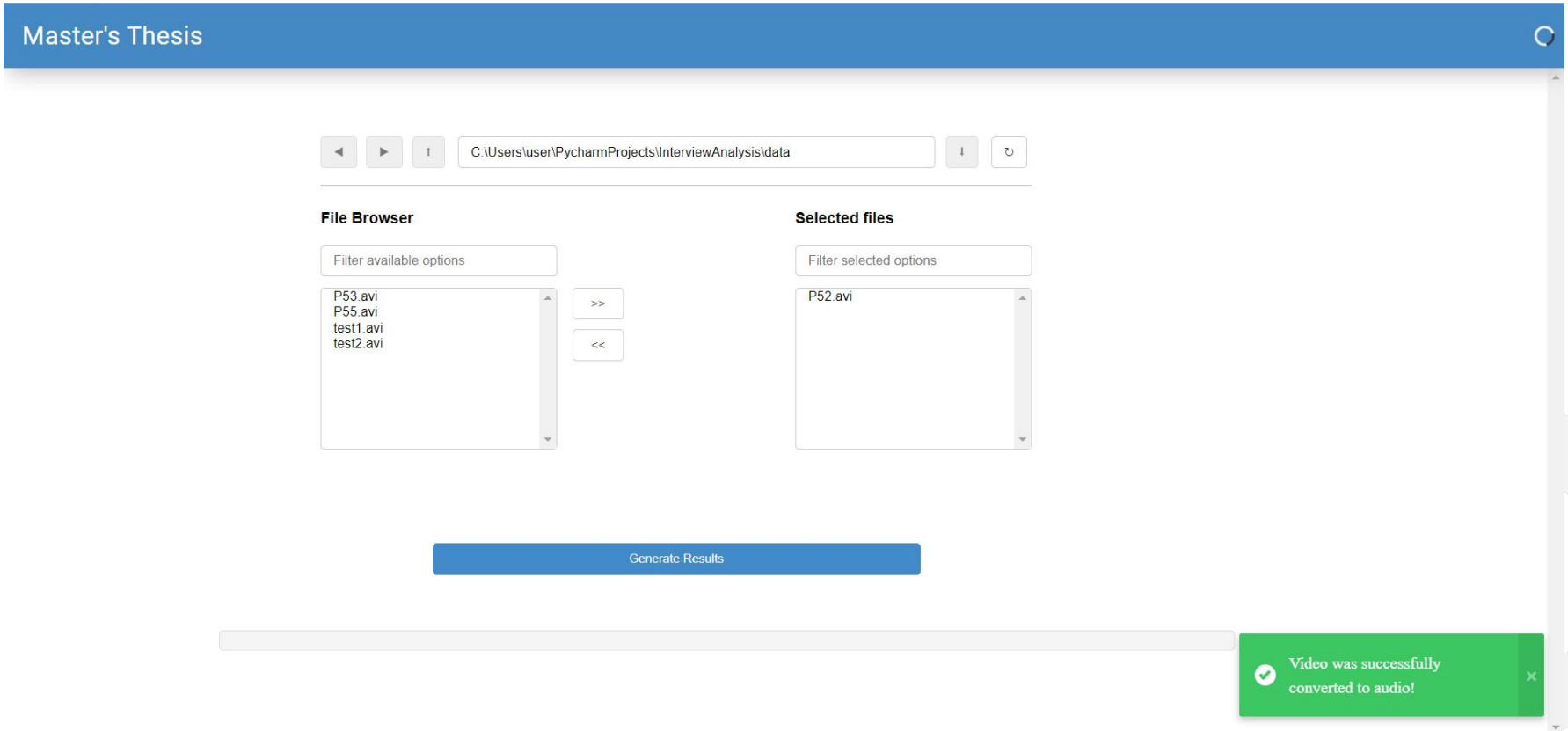
NAZARBAYEV UNIVERSITY

Adopting for UI

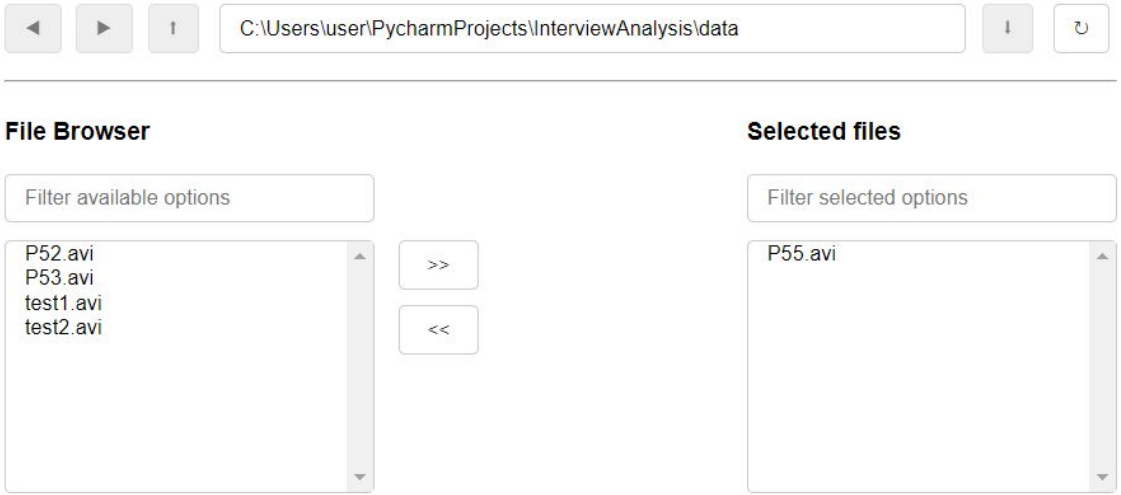


NAZARBAYEV UNIVERSITY

Demonstration



NAZARBAYEV UNIVERSITY



- Uploading the video

Final Prediction

Candidate P55 is recommended to get hired!

index	engagement	excited	eye_contact	smiling	friendly	paused	not_stressed	focused	speaking_rate	not_awkward	structured_answ
0	true	0.3	0.4	0.7	1.0	0.2	0.1	0.0	0.57093	0.9	true

- Getting results

NAZARBAYEV UNIVERSITY

Custom Dataset 1

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 4.6: Custom Dataset Results

- The custom video dataset was used to test the multimodal interview analysis model additionally, which yielded results for five videos.
- In this testing set system gives 70% accuracy on average based on self evaluation.

NAZARBAYEV UNIVERSITY

Custom Dataset 2 - Mock Interviews

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: Custom 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 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.

NAZARBAYEV UNIVERSITY

Conclusion

- The proposed multimodal interview analysis system has the potential to provide invaluable insights for hiring recommendations by utilizing deep learning techniques.
- A fusion approach is used to merge information from multiple modalities into one dataframe 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%
- 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%

NAZARBAYEV UNIVERSITY

Thank you for your attention!