Leveraging Artificial Intelligence for Emotion Recognition in Speech

Raj Shah 60004160101
Shaurya Shettigar 60004160111
Mahima Thakar 60004160120
Harshi Thaker 60004160121

Project Guide: Prof. Chetashri Bhadane

Introduction

- Human speech is a combination of linguistics and emotions and a machine is incapable of recognizing human emotions.
- Emotion detection can play a huge role in improvising the experience of the user through a healthy HCI.
- Audio Tagging is the detection and tagging of emotions within a speech sample.
- Has diverse applications.

Literature Review

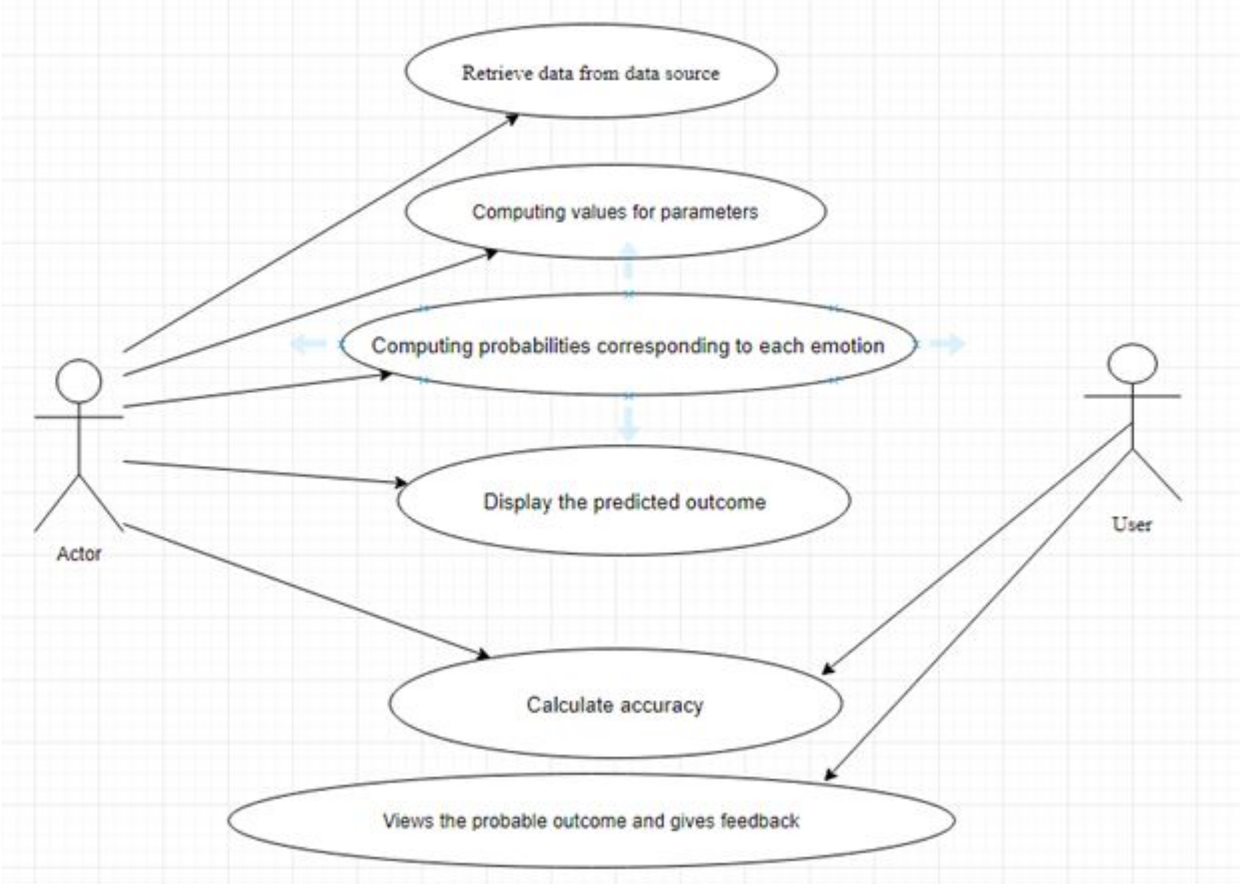
- We have referenced more than 9 papers and written a paper on it. (Paper has been accepted in Springer LNCS at the ICACTA Conference)
- This includes approaches dealing with German dataset, Berlin Dataset, Danish Dataset, etc.
- This includes implementation of algorithms like Logistic Regression, SVM, CNN, ANNs, etc

No	Features Used	Classification Method	Database and Accuracy	Classes Recognized
1	F0 maximum, F0 range, F0 mean, energy maximum, energy standard deviation, F1 maximum, F1 standard deviation, voicing rate standard deviation.	Neural Network Classifier	Berlin Database of Emotional Speech. 77.1%	Anger, Boredom, Fear, Sad, Happy, Neutral.
2	Intensity, fundamental frequency (F0), spectral contour, shimmer.voice quality, timing, eloquent, pitch, jitter, energy.	ANN, SVM(RBF), HMM	Danish emotional database, Berlin emotional database, Natural ESMBS, INTERFACE, KISMET, BabyEars,SUSAS,MPEG-4,Beihang University, etc.	Anger, Fear, Surprise, Disgust, Sad and Happy.
3	MFCC, Relative Amplitude ,LPCC , GRNN and SFS.	SVM(RBF) and Neural Network.	Berlin Database of Emotional Speech. 72%	Anger, Boredom, Disgust, Fear, Joy, Sad and Neutral

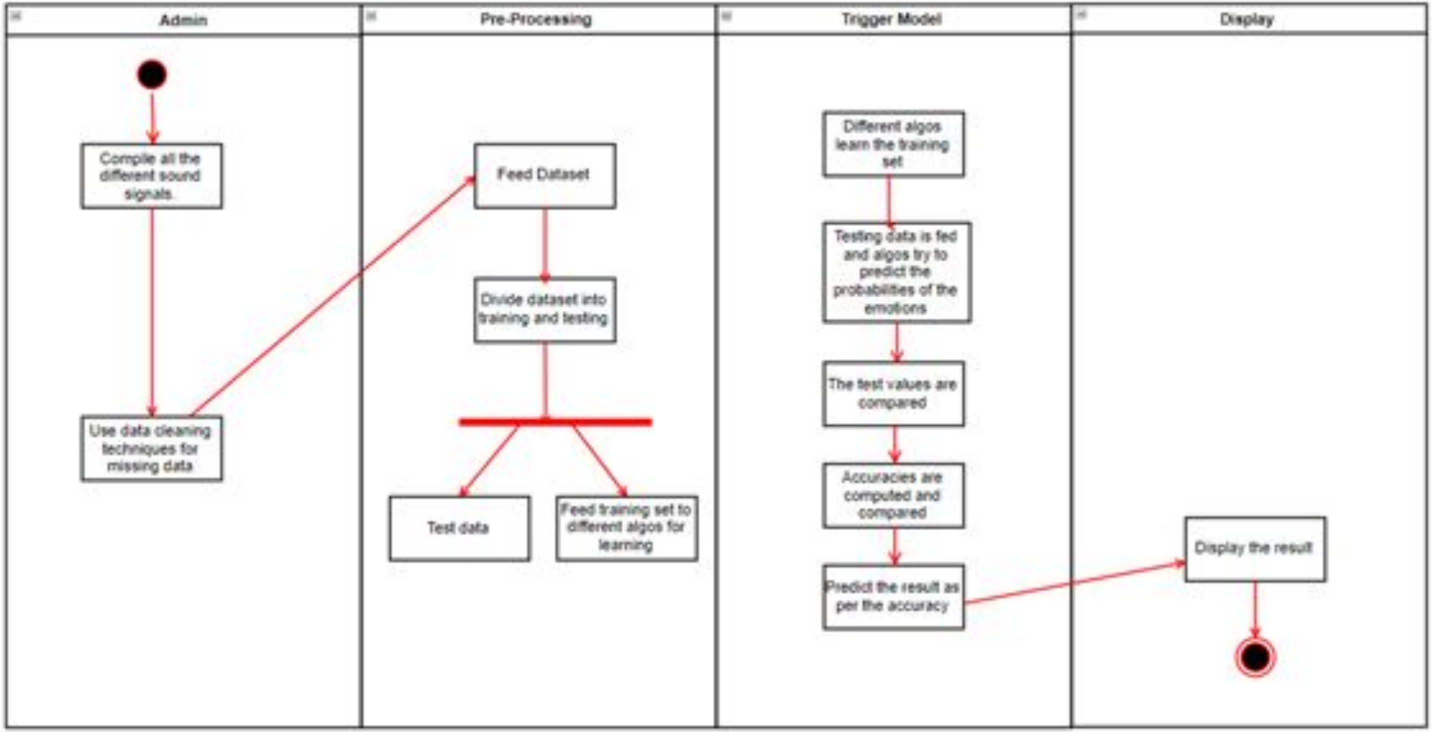
4	Tempo, amplitude's mean and maximum and pitch's maximum and deviation.	Neural Networks and SVM	Accuracy achieved was 70%	
5	Short-Time Fourier Transform, Regression of wave.	SVM, Backpropagation	German Emotional Database(68%), 4 emotion sample(91%)	Neutral, happy, sad, angry, disgust, fear, and boredom.
7	Means, Standard deviations, Maximums and Minimums of F0, Delta F0, Log energy, First and second linear prediction Cepstral coefficients(LPCC)	Enhanced Co-Training Algorithm(HMM, multi-SVM)	1800 Chinese Mandarin utterances (75.87% for Females and 80.93% for Males)	Anger, Fear, Happiness, Neutral, Sadness and Surprise
8	About 20 features including: mean, standard deviation, minimum, maximum and range, rhythm, smoothed pitch signal, etc	Maximum Likelihood Bayes (MLB), Kernel Regression (KR), KNN and CC.	1250 training utterances(Error Rate of 20.5%)	Happy, sad, anger and fear

Proposed Model

Use Case Diagram

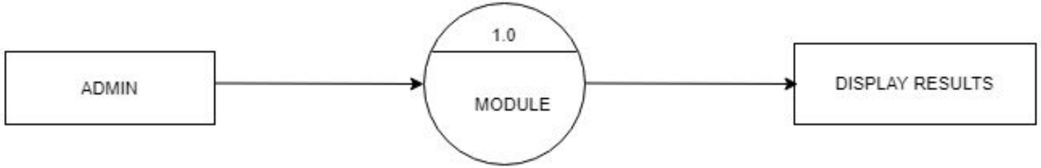


Activity Diagram

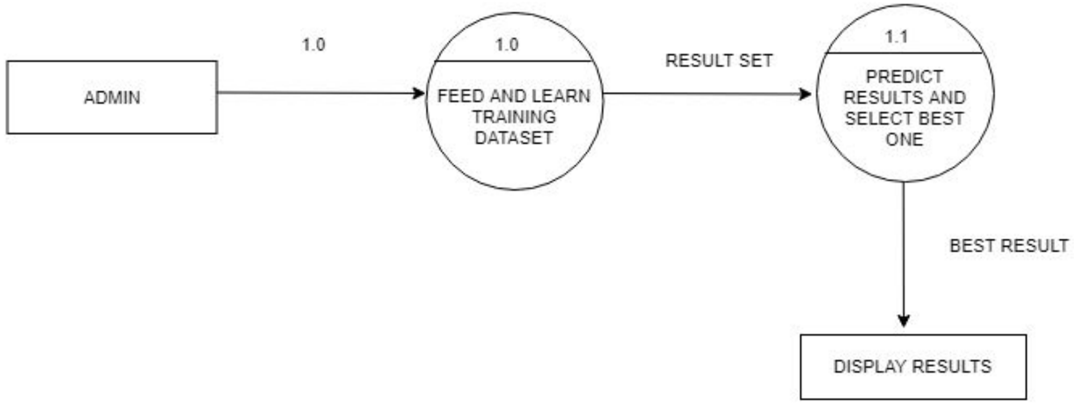


Data Modelling Diagram

LEVEL 0

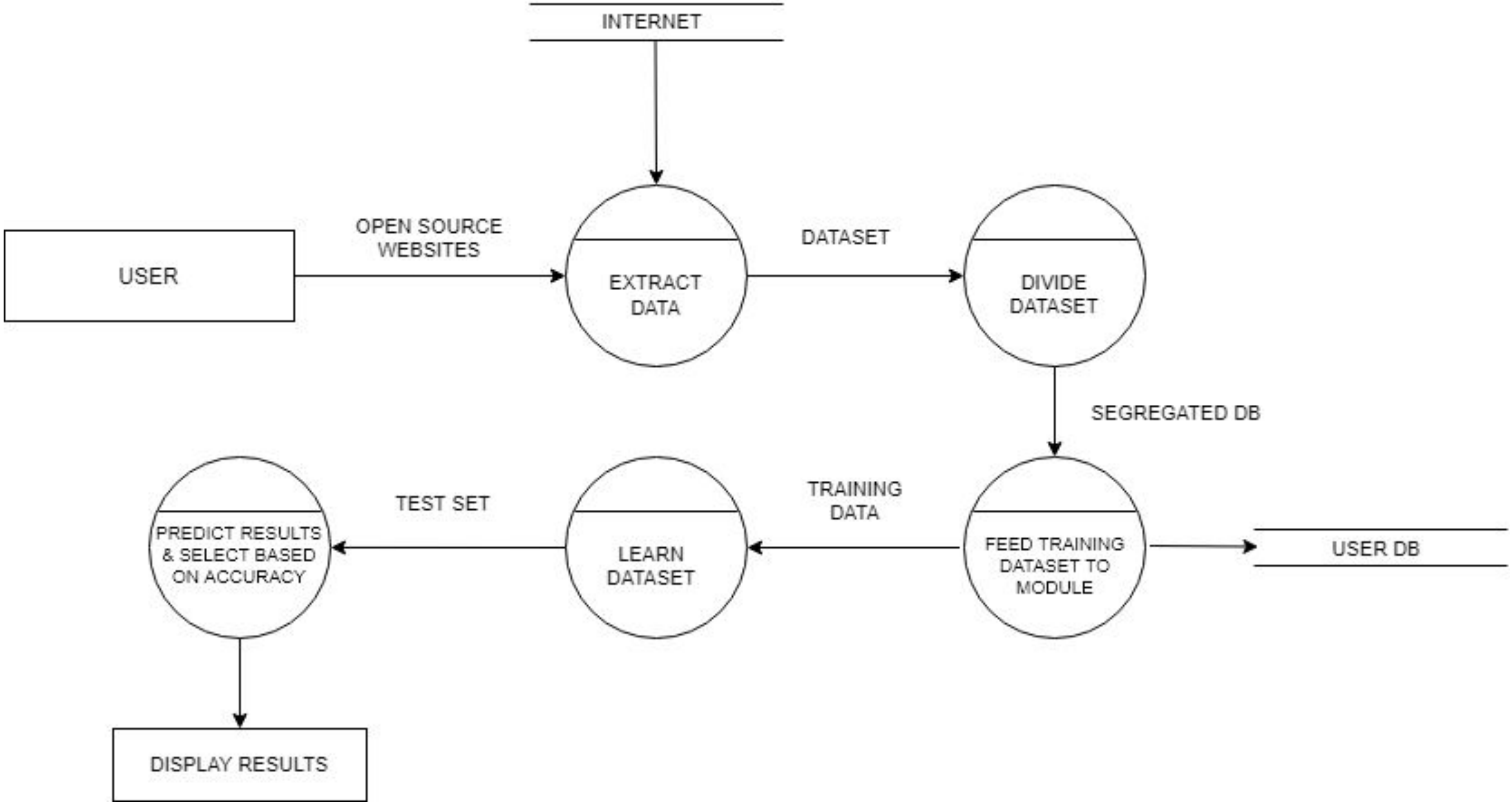


LEVEL 1

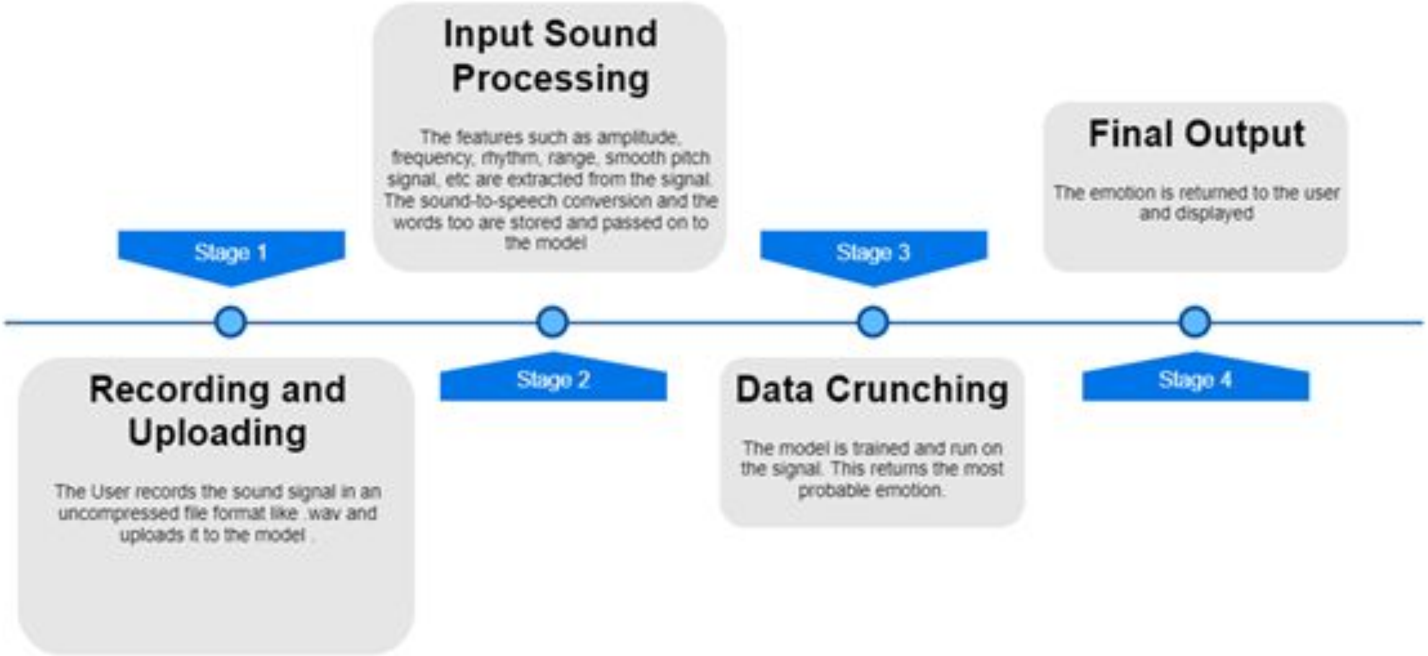

ADMIN
1.0
1.0
FEED AND LEARN TRAINING DATASET
RESULT SET
1.1
PREDICT RESULTS AND SELECT BEST ONE
BEST RESULT
DISPLAY RESULTS

LEVEL 2

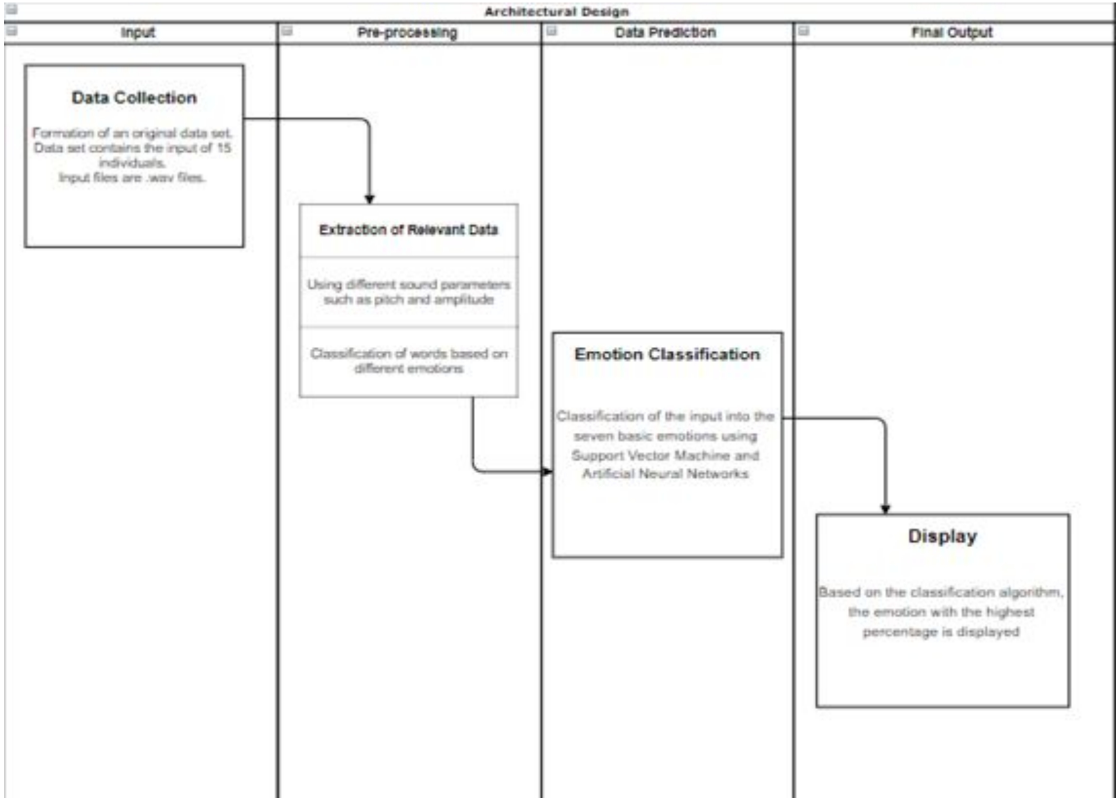
SPEECH EMOTION RECOGNITION



Functional Modelling

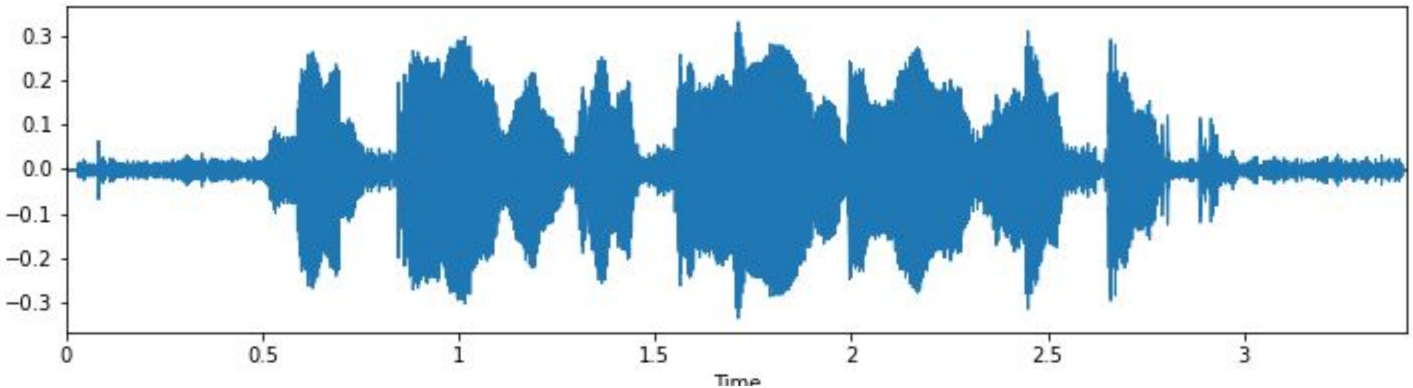


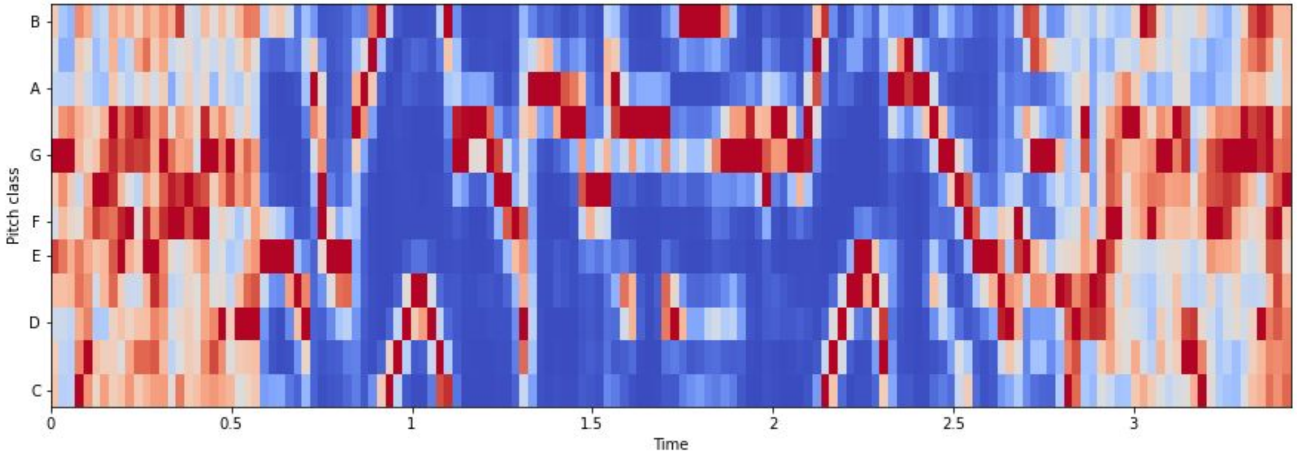
Architectural Design

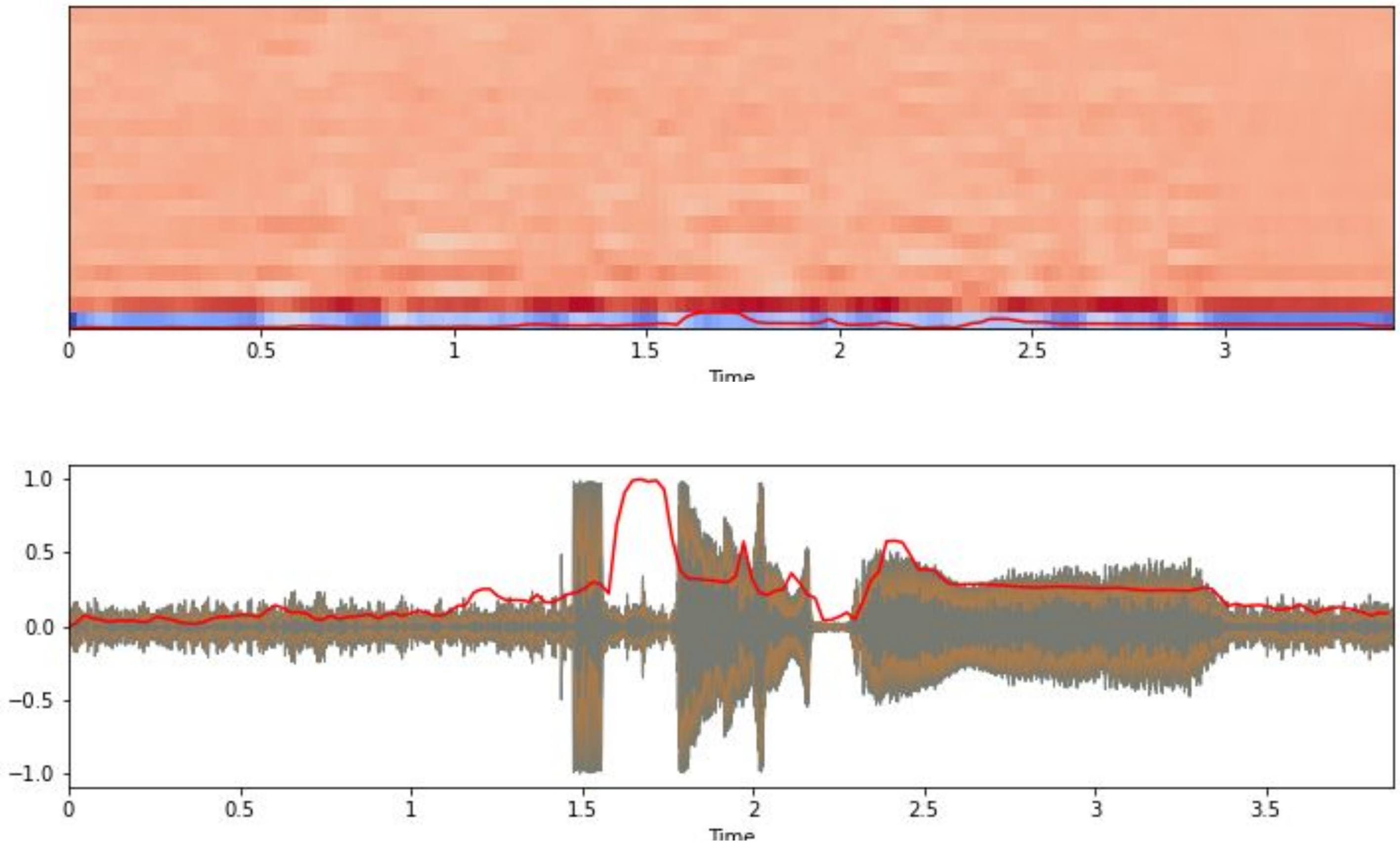


Data Collection

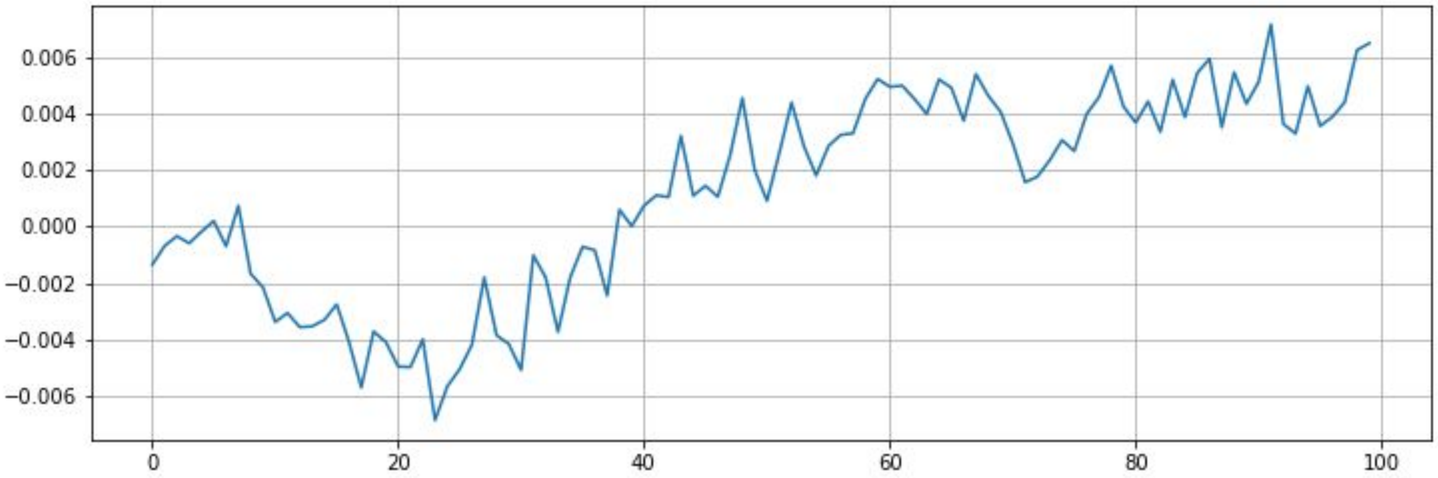
- The data set being used for our model is an original data set that we collected.
- We did not use any pre-existing datasets due to one of the following problems:-
  - Language Barrier
  - Lack of Variance
- We have used 80% percent of this data as the training set and the remaining 20% as the validation set.
- This set is formed of data inputs taken from 15 people (includes 6 Males and 8 Females, varied age range)
- The sound has been recorded in the .wav format.


0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
0.5
1
1.5
2
2.5
3
Time


B
A
G
F
E
D
C
Pitch class
0
0.5
1
1.5
2
2.5
3
Time



0
0.5
1
1.5
2
2.5
3
Time
1.0
0.5
0.0
−0.5
−1.0
0
0.5
1
1.5
2
2.5
3
3.5
Time


0.006
0.004
0.002
0.000
−0.002
−0.004
−0.006
0
20
40
60
80
100

Data Preprocessing(1/3)

- We converted the .wav input audio files into INT16 format (16 bit wav) with 16000 Hz Sampling Rate, using FFMPEG.
- And extracted many sound characteristics from them by leveraging LibROSA within the Python Audio module.
- The 20 sound characteristics included RMSE, ZCR, Mel Frequency Cepstrum Coefficients(MFCC), MFCC_Delta, Tempo, Loudness, Gender, Pitch, Chroma, Beats, Contrast, RollOff, Tonnetz, Harmonic, Percussion, etc.
- Certain characteristics such as MFCC, had more than 7000 elements in their array; The mean was considered.

df - DataFrame

Index	ID	SONG_NAME	rmse	zcr	mfcc	mfcc_delta	loudness	tempo	chroma_stft_mean	chroma_cq_mean	beats	chroma_cens_mean	mel_mean	cent_mean	spec_bw_mean	contrast_mean	rolloff_mean	poly_features	tonnetz	harm_mean	perc_mean	class
0	1	Aagam Shah_Anger_1.wav	0.0748	0.1	-20.3	-0.000678	-14.4	136	0.433	0.552	962	0.269	3.39	1.78e+03	1.7e+03	24.4	3.47e+03	0.852	-0.00541	-7.29e-06	-2.46e-05	0
1	2	Aagam Shah_Anger_2.wav	0.0995	0.142	-18.7	0.00101	-14.2	86.1	0.422	0.549	224	0.278	4.23	2.29e+03	1.87e+03	24.3	4.35e+03	1.26	-0.00672	-5.24e-06	-3.62e-05	0
2	3	Aagam Shah_Anger_3.wav	0.0956	0.113	-20.4	-0.00465	-13.6	144	0.4	0.521	566	0.272	4.47	1.96e+03	1.81e+03	24.6	3.91e+03	1.11	0.00126	-6.52e-06	-1.38e-05	0
3	4	Aagam Shah_Anger_4.wav	0.0634	0.0793	-20.7	-0.00163	-15.9	108	0.412	0.545	358	0.274	2.55	1.47e+03	1.43e+03	23.6	2.62e+03	0.678	-0.00118	-6.02e-06	-1.63e-05	0
4	5	Aagam Shah_Anger_5.wav	0.0992	0.082	-19.1	2.63e-05	-13.9	152	0.402	0.516	175	0.27	5.01	1.63e+03	1.58e+03	24.5	3.05e+03	1.07	-0.0062	-5.67e-06	-2.78e-05	0
5	6	Aagam Shah_Anger_6.wav	0.0663	0.0846	-21.8	0.00499	-14.1	68	0.435	0.559	843	0.274	3.01	1.58e+03	1.57e+03	23.6	3.05e+03	0.747	-5.87e-05	-8.52e-06	-2.85e-05	0
6	7	Aagam Shah_Anger_7.wav	0.0804	0.108	-20	0.0012	-13.2	103	0.461	0.575	310	0.272	4.2	1.79e+03	1.55e+03	23.3	3.45e+03	0.904	-0.00405	-5.85e-06	-2.78e-05	0
7	8	Aashreen_Anger-1.wav	0.156	0.0645	-5.42	0.0068	-14.4	136	0.343	0.471	466	0.263	8.24	1.53e+03	1.91e+03	21.3	2.89e+03	1.26	-0.00439	-4.14e-07	-9e-07	0
8	9	Aashreen_Anger-2.wav	0.1	0.0937	-6.61	0.0297	-17.4	95.7	0.381	0.501	354	0.258	3.99	1.82e+03	2.01e+03	20.5	3.54e+03	0.816	-0.00136	3.37e-06	-1.32e-06	0
9	10	Aashreen_Anger-3.wav	0.145	0.0734	-5.7	0.0226	-14.7	172	0.328	0.444	677	0.255	7.43	1.76e+03	2.08e+03	20.9	3.42e+03	1.09	0.00883	-1.23e-06	-3.91e-06	0
10	11	Aashreen_Anger-4.wav	0.126	0.0554	-4.81	0.0149	-16.2	83.4	0.367	0.481	162	0.259	5.77	1.51e+03	2.03e+03	20.2	2.99e+03	1.01	0.00122	1.54e-05	-9.19e-06	0
11	12	Aashreen_Anger-5.wav	0.105	0.0454	-5.29	-1.11e-05	-15.9	108	0.421	0.481	244	0.249	4.76	1.38e+03	1.96e+03	20.1	2.79e+03	0.798	0.0134	-1.21e-05	-6.61e-06	0
12	13	Aashreen_Anger-6.wav	0.0847	0.0775	-5.2	0.022	-18.7	89.1	0.4	0.53	307	0.268	3.02	1.75e+03	2.09e+03	20.2	3.35e+03	0.728	0.00252	-6.31e-06	-1.29e-05	0
13	14	Aashreen_Anger-7.wav	0.106	0.0984	-5.02	0.0219	-16.5	161	0.411	0.502	342	0.259	4.68	1.96e+03	2.1e+03	20.2	3.65e+03	0.873	0.00116	7.46e-06	-9.96e-06	0
14	15	Aditi Chavan_ANGER 1.wav	0.0596	0.0509	-8.36	0.0615	-17.8	129	0.388	0.462	335	0.265	1.22	1.23e+03	1.54e+03	21.3	2.34e+03	0.532	0.0085	-2.2e-07	-3.87e-06	0
15	16	Aditi Chavan_ANGER 2.wav	0.0503	0.0874	-8.48	0.0602	-19.7	161	0.394	0.517	333	0.272	0.891	1.8e+03	1.84e+03	20.3	3.46e+03	0.443	0.00514	2.23e-07	-4.85e-06	0
16	17	Aditi Chavan_ANGER 3.wav	0.0539	0.0731	-8.6	0.0508	-17.7	117	0.378	0.489	252	0.275	0.952	1.49e+03	1.6e+03	20.7	2.72e+03	0.523	0.00434	-1.34e-07	-2.39e-07	0
17	18	Aditi Chavan_ANGER 4.wav	0.0512	0.0543	-8.6	0.0561	-17.9	123	0.358	0.5	225	0.271	0.883	1.37e+03	1.6e+03	20.7	2.69e+03	0.495	0.00395	1.17e-07	1.01e-05	0
18	19	Aditi Chavan_ANGER 5.wav	0.065	0.0406	-8.19	0.0384	-16.6	152	0.388	0.491	269	0.269	1.46	1.07e+03	1.41e+03	20.8	2.03e+03	0.578	0.00437	1e-07	-2.4e-05	0
19	20	Aditi Chavan_ANGER 6.wav	0.0557	0.0818	-7.88	0.0482	-17.5	92.3	0.435	0.531	185	0.278	1.1	1.51e+03	1.61e+03	20.4	2.75e+03	0.564	-0.00106	-3.18e-07	-1.14e-05	0
20	21	Aditi Chavan_ANGER 7.wav	0.0618	0.112	-8.17	0.056	-17	161	0.395	0.506	482	0.263	1.29	1.87e+03	1.86e+03	20.7	3.5e+03	0.561	-0.000643	-2.39e-08	-1.55e-05	0
21	22	Dhrashti_Angry_1.wav	0.0245	0.0628	-12.1	0.0661	-30.4	123	0.354	0.461	384	0.256	0.197	1.4e+03	1.77e+03	22.1	2.8e+03	0.2	-0.00948	-0.000239	-2.6e-05	0
22	23	Dhrashti_Angry_2.wav	0.0172	0.0979	-13.3	0.0472	-32.5	144	0.368	0.453	383	0.258	0.115	1.87e+03	1.95e+03	21.4	3.81e+03	0.154	-0.00469	-6.03e-06	-7.67e-05	0
23	24	Dhrashti_Angry_3.wav	0.0133	0.114	-12.8	0.0521	-33.9	185	0.419	0.524	224	0.267	0.0696	1.96e+03	1.84e+03	20.4	4.07e+03	0.141	0.00678	-3.08e-05	-6.37e-05	0
24	25	Dhrashti_Angry_4.wav	0.0151	0.114	-12.8	0.041	-34.1	185	0.4	0.527	473	0.276	0.0857	2.15e+03	2.21e+03	21	4.83e+03	0.158	-0.00166	4.59e-06	-5.91e-05	0
25	26	Dhrashti_Angry_5.wav	0.0143	0.0582	-14.2	0.0695	-33.9	185	0.427	0.505	193	0.252	0.0853	1.38e+03	1.76e+03	21	2.84e+03	0.12	0.00866	2.59e-05	-4.2e-05	0
26	27	Dhrashti_Angry_6.wav	0.0118	0.0928	-13.3	0.0524	-35.6	117	0.414	0.508	320	0.258	0.059	1.8e+03	1.9e+03	20.8	3.63e+03	0.12	0.00144	1.2e-05	-5.37e-05	0
27	28	Dhrashti_Angry_7.wav	0.00685	0.114	-13.1	0.0316	-40.5	78.3	0.491	0.591	195	0.271	0.016	2e+03	1.95e+03	20	4.02e+03	0.0835	-0.00124	3.27e-05	-6.29e-05	0
28	29	Dhwani_Anger-1.wav	0.112	0.0703	-3.64	0.0182	-15.6	144	0.378	0.484	306	0.26	4.24	1.91e+03	2.12e+03	20.8	3.78e+03	1.09	0.000802	3.97e-07	-3.91e-06	0
29	30	Dhwani_Anger-2.wav	0.132	0.098	-1.41	-0.00452	-14.4	144	0.374	0.514	403	0.271	5.69	2.03e+03	2.08e+03	20.3	3.82e+03	1.18	0.00177	-1.93e-05	-3.62e-06	0
30	31	Dhwani_Anger-3.wav	0.107	0.0745	-3.92	-0.00455	-16.8	112	0.316	0.479	250	0.264	3.6	1.95e+03	2.07e+03	20.7	3.74e+03	1	0.0121	-8.74e-06	-8.09e-06	0

Format Resize Background color Column min/max Save and Close Close

Type here to search ENG IN 06:43 PM 07/May/2020

Data Preprocessing(2/3)

- **Label Encoding** refers to converting the labels into numeric form so as to convert it into the machine-readable form.
  - This transformer should be used to encode target values, *i.e.* `y`, and not the input `X`.

- For each value in a feature, **MinMaxScaler** subtracts the minimum value in the feature and then divides by the range.
  - The range is the difference between the original maximum and original minimum.
  - The default range for the feature returned by MinMaxScaler is 0 to 1.

Data Preprocessing(3/3)

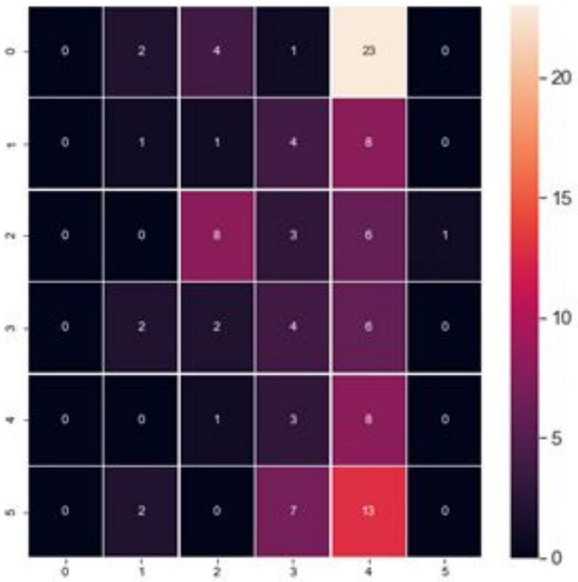
- There were certain Features such as pitch, gender and the speech words that we did removed from the data frame.
- We removed pitch and gender because they were highly correlated with other features and did not better our emotion recognition model.
- We did implement the speechtotext module in our project, however, adding comprehensive sentiment analysis modules were pushing our project out of scope.
- This became our **Feature Set**, on which we applied the classifiers.

Implementation

We have tried using multiple algorithms for the classification task with varying results:-

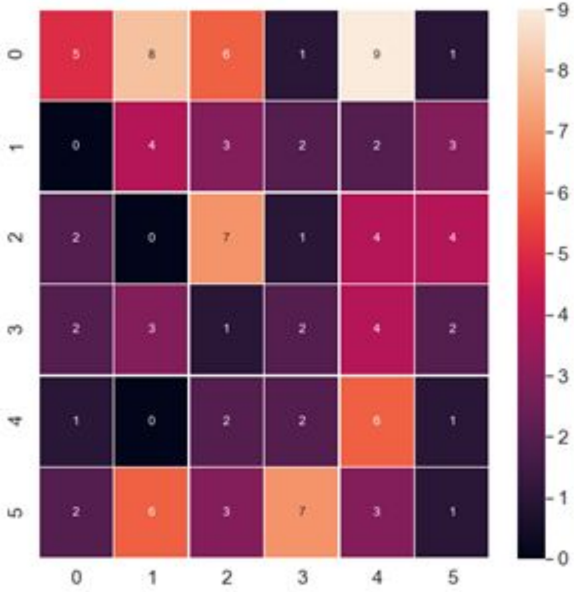
1. Naïve Bayes
2. Logistic Regression
3. Gradient Boosting
4. Random Forest
5. Support Vector Machine
6. Artificial Neural Network with Grid Search.

Naive Bayes Algorithm HeatMap



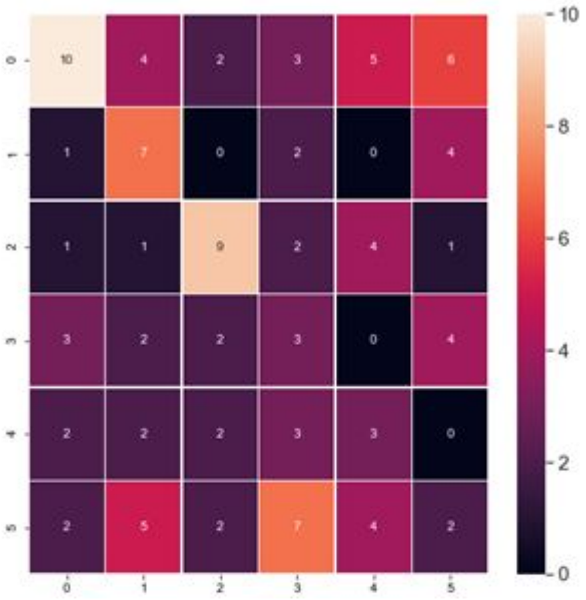
Accuracy: 19%

Logistic Regression Algorithm HeatMap



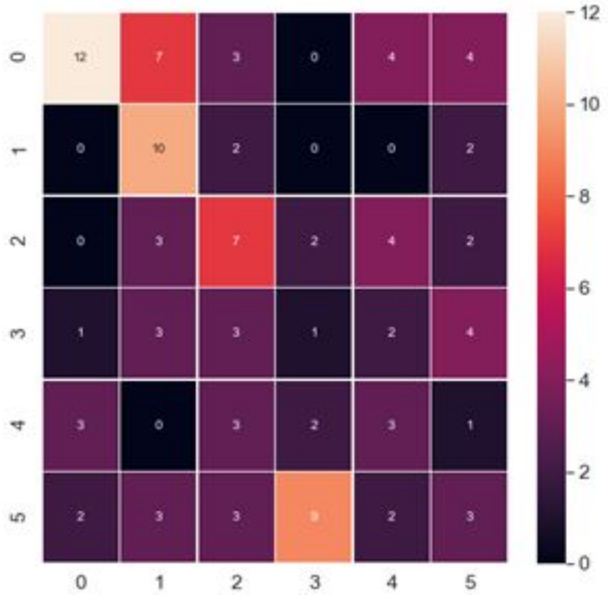
Accuracy: 22.72%

Support Vector Machine HeatMap



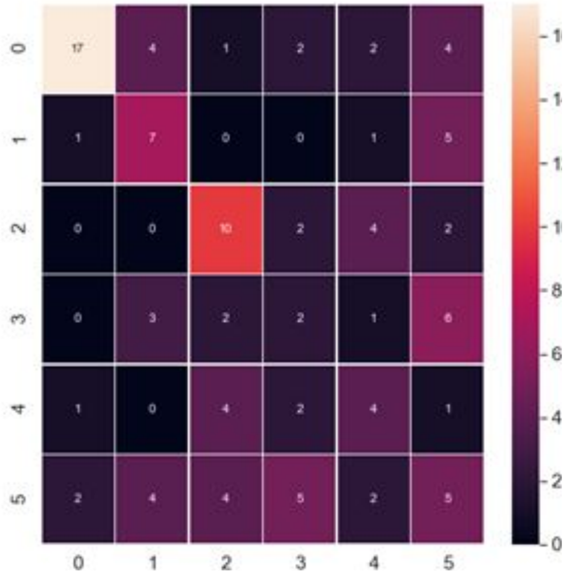
Accuracy: 31%

Random Forest Algorithm HeatMap

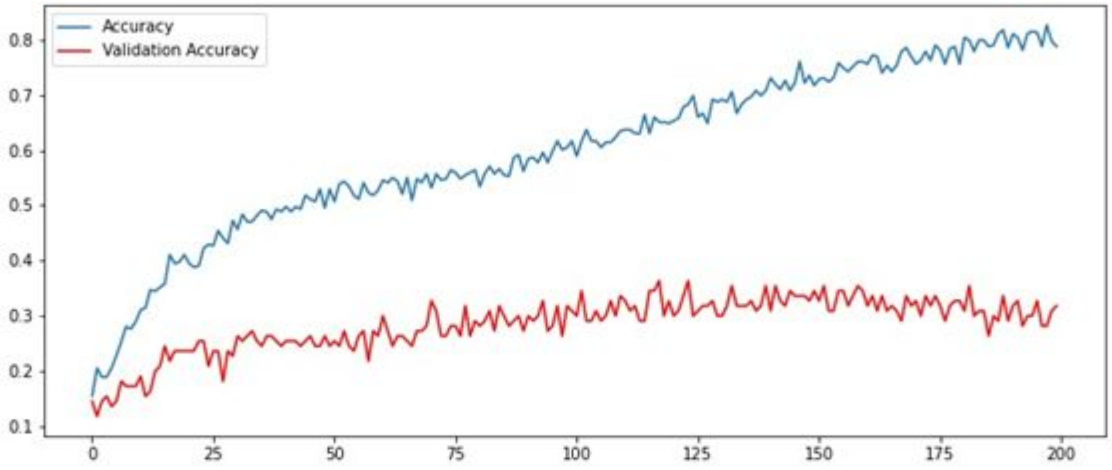


Accuracy: 32.72%

Gradient Boosting HeatMap



Artificial Neural Network with GridSearch



Training Accuracy: 43.02%
Validation Accuracy: 41.99%

Training Accuracy: 81%
Validation Accuracy: 38%

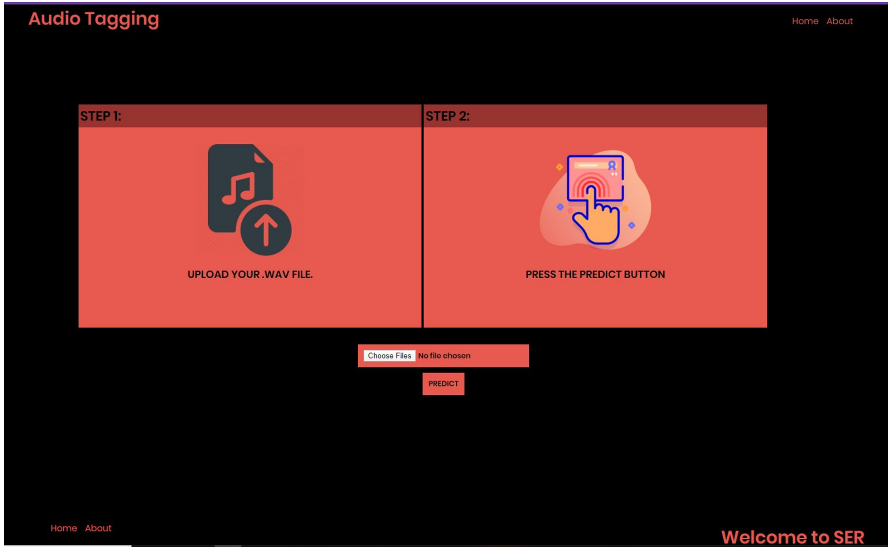
Why Gradient Boosting and not ANN?

- According to our own literature reviews, SVM and ANN should have had the greatest accuracies.
- ANN did have the greatest training accuracy of 81%, however it just had a validation accuracy of 38%. Indicating some amount of overfitting.
- Gradient Boosting has the 2nd highest result out of the other algorithms, thus we have gone forward with implementing that.
- One of our hypothesis for the underperformance of ANN, is the lack of data.
- Despite having more than 600 Training Samples, ANN requires far more to avoid overfitting.

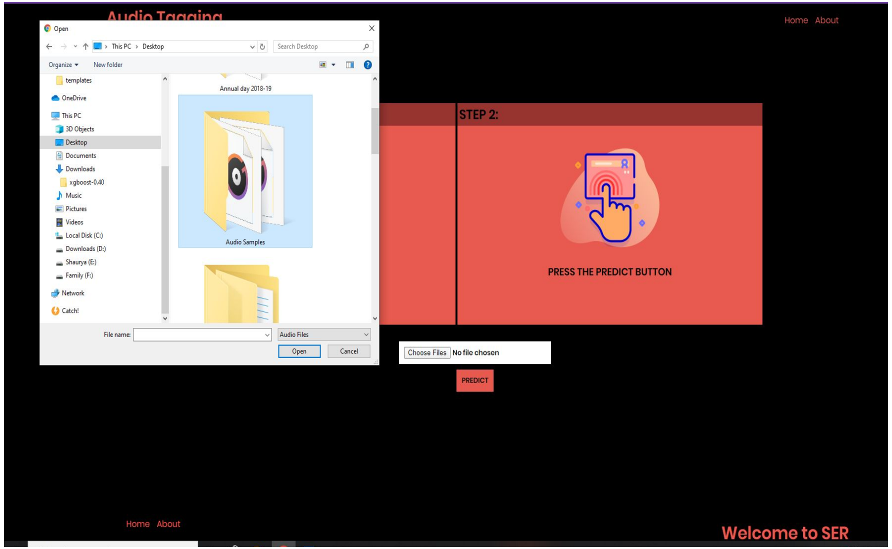
Side Note: Using Pickle

- The pickle module serializes objects so they can be saved to a file, and loaded in a program again later on.
- It is the conversion of an object from *data in RAM* to *text on disc*.
- This is because you can save them to be able to make new predictions without training the model all over again.
- This drastically reduced the runtime so that later on our project can handle and interact with speech samples more dynamically.

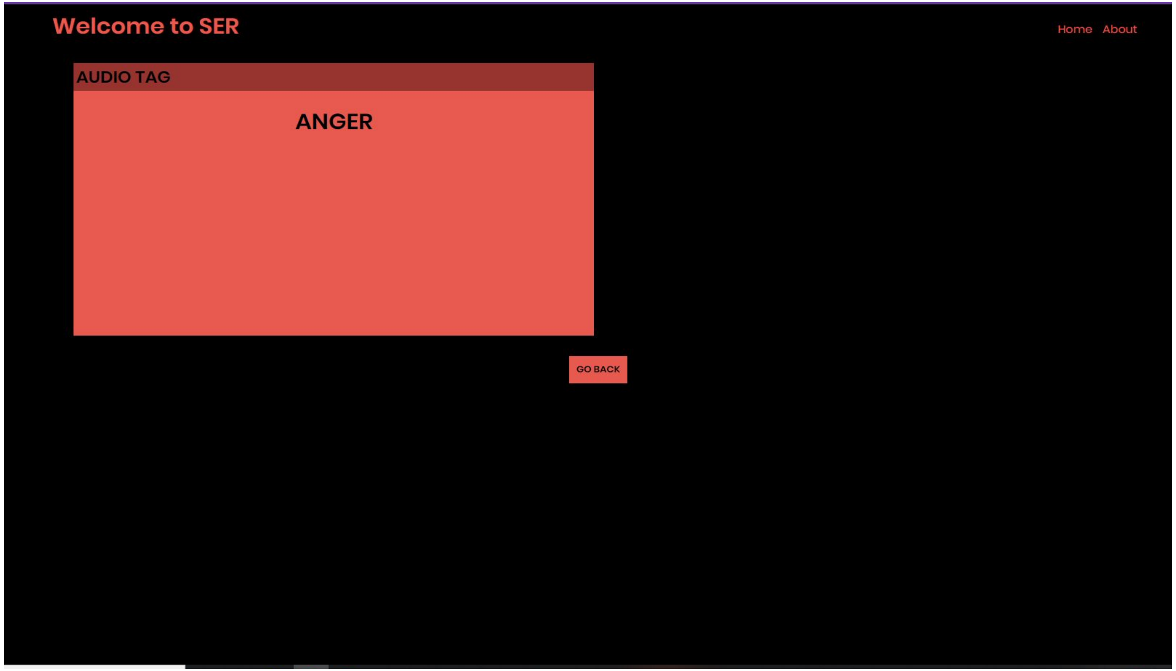
Working of the project with GUI



Home page



Choosing the audio file to upload



Result display after clicking the predict button

Results and Discussions

EmotiW 13	AFEW 3.0	US	film	~.8/315/1088	7	clip	1582 oS	SVM	.2244 WA
EmotiW 14	AFEW 4.0	US	film	~1.0/428/1368	7	clip	1582 oS	SVM	.2678 WA
MEC 16	CHEAVD	CN	film/TV	2.3/238/2852	8	clip	88 oS	RF	.2402 MAP/.2436 WA
MEC 17	CHEAVD 2.0	CN	film/TV	7.9/527/7030	8	clip	88 oS	SVM	.392 MAP/.405 WA

- A 48% accuracy may seem disappointing, however it beats the state of the art SER Engines.
- Especially those who classify only with an audio file as an Input.
- What about Data Augmentation?

Conclusions and Future Scope

- Implement a module for Facial recognition.
- Take an input for gender and classify differently for both genders.
- Implement sentiment analysis to make the SpeechToText module relevant.
- Dynamic usage.