Generalizing Speech Emotion Recognition: A Multi-Model and Multi-Dataset Approach

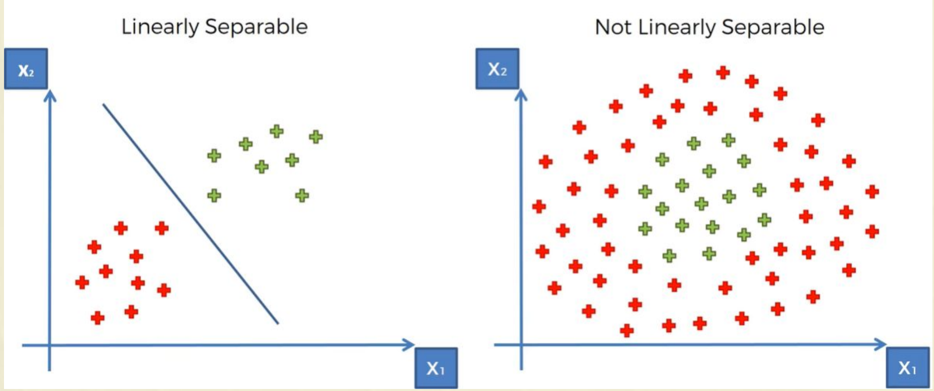
Group 15:
Sweta Patel (100915164)
Vallika Kasibhatla (100928820)

Date: 10th June
ENGR 5510G: Foundations of Software Engineering
Instructor: Dr. Sanaa Alwidian

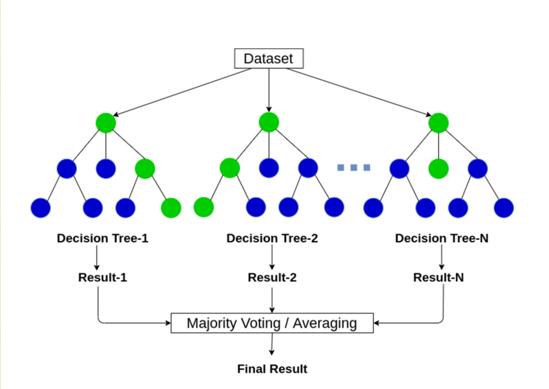
TABLE OF CONTENTS

KEYWORDS

Speech Emotion Recognition	Identifying and Analyzing the emotions through the voice of the speaker
Support Vector Machine (SVM)	It is a model that generates a decision boundary
Random Forest (RF)	It operates by constructing a multitude of decision trees
Neural Networks	It is a machine learning model that learns complex patterns by making neuron structures



SVM (a) linear (b) non-linear



Random Forest

01

INTRODUCTION

Emotion Detection Through Speech

Emotion Detection has become an emerging research subject.

- ➔ Human emotions are challenging to guess
- ➔ Analyzing an audio note has high complexity
- ➔ Understanding sentiments of humans can improve productivity and efficiency

02

RESEARCH GAPS AND PROPOSED SOLUTION

Research

Ref	Datasets Used	Pre-Processing and Feature Extraction	Model	Results Of the Papers
[2]	IEMOCAP	MEL coefficients	DNN, CNN, LSTM	64.78%
[3]	EEG Signals	Feature fusion from Deep Belief Network and Electro-Dermal Activity (EDA), Photoplethysmogram (PPG), and Zygomaticus Electromyography (EMG) sensors.	Fine Gaussion SVM	'Happy' Emotion-100% 'Neutral' Emotion-53%
[4]	RAVEDESS	MFCCs, MEL Spectrograms, Chroma, Tonnetz	MLP, SVM	MLP-79% SVM-72%
[5]	TESS	MFCC, MEL Spectrograms, Chroma, Tonnetz	KNN, Decision Trees, Extra Tree Classifier	KNN-98% Extra Tree Classifier-99% Decision Trees-92%
[6]	RAVEDESS, TESS	MFCCs	CNN, AlexNet, ResNet50	RAVEDESS-80% with CNN TESS- 96% CNN
[7]	EMODB RAVEDESS TESS IEMOCAP	STFT, Feature extracted using Multi-Time Scale (MTS) versions of learned kernel	CNN	EMODB-70.97% RAVEDESS-55.85% TESS-53.05% IEMOCAP-55.01%
[8]	IEMOCAP, MELD	Low-Level Descriptors BoAW	RNN	IEMOCAP 60.87%
[9]	Berlinemodb	Spectrogram Generation FFT	CNN, AlexNet model	84.3%
[10]	SAVEE	MFCC, MSF	CNN	83.61%
[11]	RAVDESS, TESS	MFCC STFT	SVM, HMM, ELM, LSTM, CNN	CNN - 85%
[12]	RAVDESS	MFCC	LSTM	84.81%
[13]	EmoDB RED	VAD MFCC SBS	SVM, LDA	EmoDB: LDA: 78%, SVM: 80% RED: LDA: 71%, SVM: 73%

Research Gaps

- Having speakers only of English descent, can create a bias while training and testing
- Using single datasets while training
- Mixed emotions or rapid emotions switching wouldn't be detected.
- Speaker diarization
- Papers have highlighted misclassifications but havent explored them
- The classification models used are highly focused on neural networks, maybe it's too complex.
- Difficulty in finding the best feature extraction steps

Proposed Solution

Generalizability by including through analyzing

Focusing on Simpler Classification Models

Ensemble Learning

03

DATASETS

DATASETS

01
RAVDESS
840 audio files
(8 emotions: Angry, Surprised, Disgust, Neutral, Happiness, Sadness, Calm)

02
CREMA-D
7442 audio files
(6 emotions: Anger, Disgust, Fear, Happy, Neutral, and Sad)

03
SAVEE
480 audio files
(7 emotions: Anger, Disgust, Fear, Happiness, Sad, Surprise, Neutral)

04
TESS
2800 audio files
Canadian Speaker
(7 emotions: Sad, Surprise, Angry, Neutral, Happy, Fear, Disgust)

05
EMODB
535 audio files
German Speakers
(7 emotions: Angry, boredom, disgust, neutral, fear, happiness, sadness)

12097 Audio Files
Anger, Happy, Sad, Fear, Disgust and Neutral

04

DATA PREPROCESSING

Steps

1

	AudioPath	Label
	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	calm
3	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	neutral
4	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	happy
5	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	sad
6	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	neutral
7	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	calm
8	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	sad
9	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	neutral
10	/content/drive/MyDrive/SoftwareAudio/RAVEDESS/Actor_20/03-01-(	happy

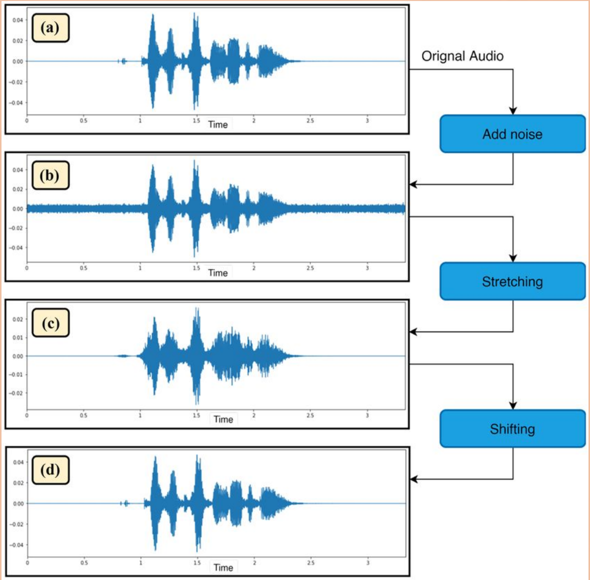
Before Extraction: Audio files are stored as raw files on disk, with their paths and labels recorded in a DataFrame.

After Extraction: Features are extracted and stored in NumPy arrays, which can be used directly for model training or further analysis.

2

Data Augmentation:

- **<u>Noise Addition</u>**
- **<u>Time Stretching</u>**
- Pitch
- Shifting



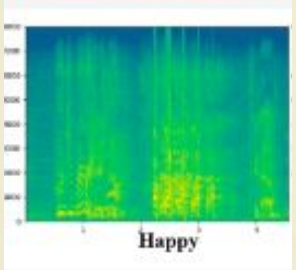
Feature extraction

3

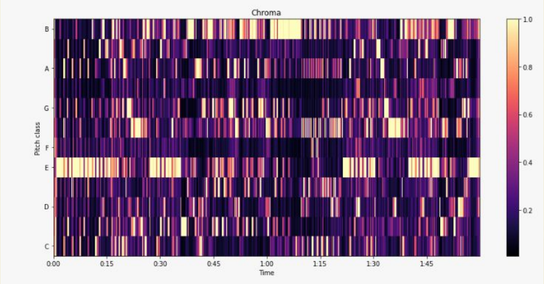
Feature Extraction:

- Zero Crossing Rate
- Chroma Feature
- MFCC
- Root Mean Square Energy
- Spectral Rolloff, Centroid, Contrast, Bandwidth
- Tonnets

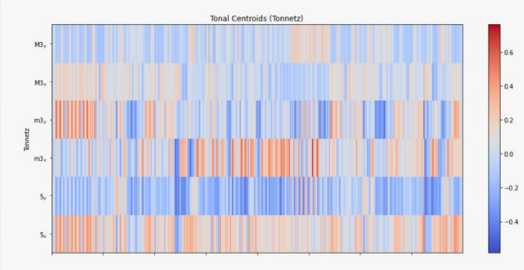
1	2	3	4	5	Label
0.017605	-0.03897	-0.05874	0.016288	0.010934	happy
0.0065	-0.0394	0.009756	0.000769	0.003263	happy
0.015879	-0.04392	-0.05527	0.019473	0.000833	happy
0.034252	-0.05858	-0.05576	0.014269	-0.00525	neutral
0.008455	-0.0629	-0.01463	0.000269	-0.00231	neutral
0.036998	-0.0514	-0.04058	0.007093	-0.0107	neutral
0.000923	-0.00513	0.039073	-0.00357	0.029264	sad
0.000713	-0.01462	0.044512	-0.00361	0.018366	sad
0.009619	-0.02456	0.034347	0.005985	0.027446	sad
0.019868	-0.01268	-0.0049	0.001143	0.013344	calm
0.019661	-0.00816	0.024284	0.001049	0.011365	calm
0.02126	-0.02695	-0.00691	-0.00487	0.002568	calm



MFCC



Chroma



Tonnets

05

MODEL BUILDING

COMPARISONS

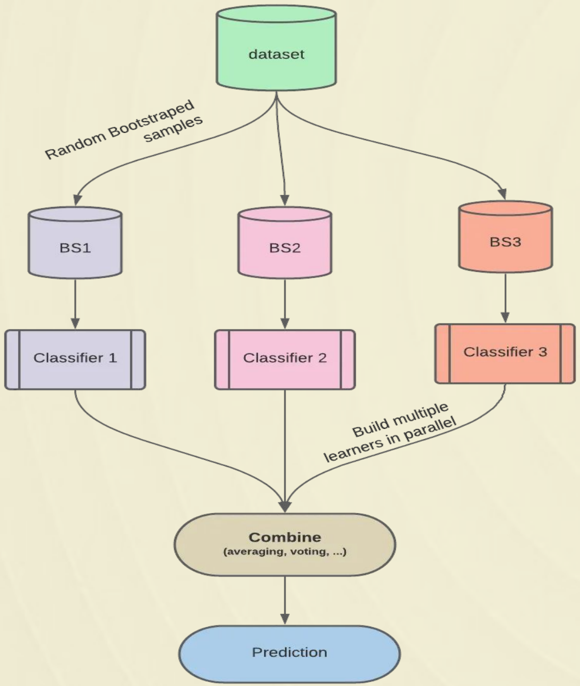
Random Forest	Support Vector Machines
• Generally faster to train and predict with large datasets • Can handle missing values and outliers more • With a large number of trees can be complex • Robust to overfitting • Less sensitive to hyperparameters	• Computationally intensive and slow • SVMs do not handle missing values and outliers well • Result in a single decision boundary so simpler to interpret • Reduces overfitting, especially in high-dimensional spaces. • Sensitive to hyperparameters

Improved Generalization
Enhanced Performance
Reduced Variance and Bias
Robustness to Different Data Characteristics

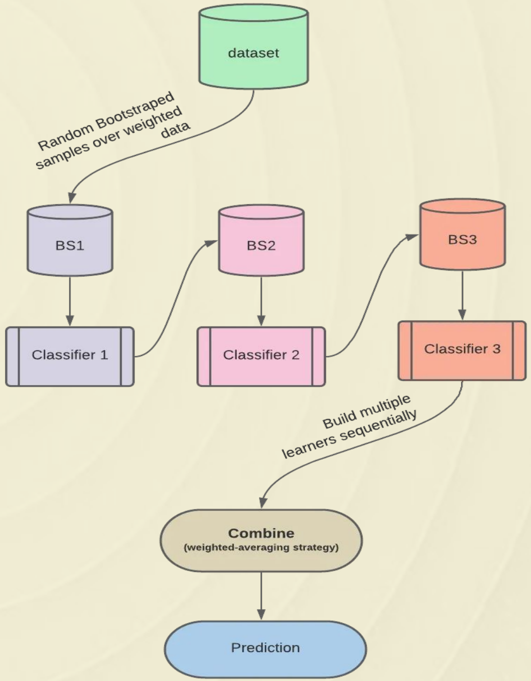
Ensemble Learning

To aggregate multiple models to obtain a combined model that outperforms every single model in it.

Bagging



Boosting



Results on Individual Datasets

RAVDESS

Accuracy (Boosted SVM and RF): 0.846064814814

	precision	recall	f1-score
angry	0.87	0.91	0.89
calm	0.79	0.94	0.86
disgust	0.84	0.84	0.84
fearful	0.88	0.83	0.86
happy	0.81	0.84	0.82
neutral	0.83	0.78	0.80
sad	0.83	0.73	0.78
surprised	0.90	0.88	0.89
accuracy			0.85
macro avg	0.85	0.84	0.84
weighted avg	0.85	0.85	0.85

CREMA-D

Accuracy (Boosted SVM and RF): 0.733766233766

	precision	recall	f1-score
angry	0.74	0.88	0.80
disgust	0.74	0.66	0.70
fear	0.78	0.61	0.69
happy	0.73	0.72	0.72
neutral	0.71	0.70	0.71
sad	0.70	0.82	0.76
accuracy			0.73
macro avg	0.74	0.73	0.73
weighted avg	0.74	0.73	0.73

SAVEE

Accuracy (Boosted SVM and RF): 0.857638888888

	precision	recall	f1-score
anger	0.82	0.72	0.77
disgust	0.86	0.84	0.85
fear	0.77	0.83	0.80
happiness	0.65	0.83	0.73
neutral	0.94	0.99	0.96
sadness	0.92	0.92	0.92
surprise	0.93	0.72	0.81
accuracy			0.86
macro avg	0.84	0.83	0.83
weighted avg	0.86	0.86	0.86

TESS

Accuracy (Boosted SVM and RF): 0.994444444444

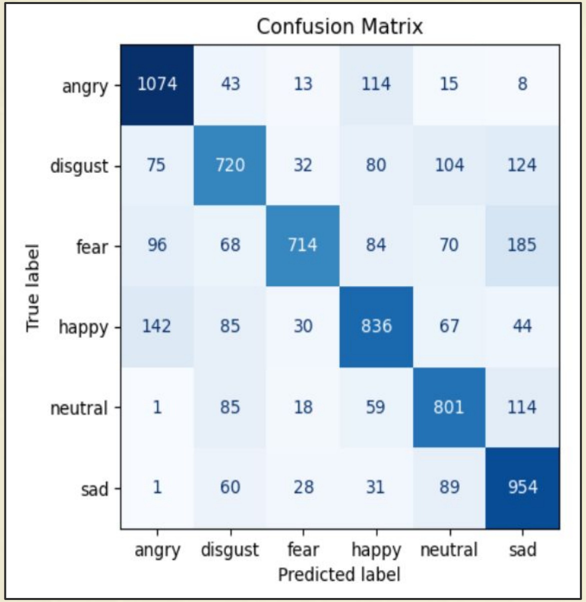
	precision	recall	f1-score
angry	1.00	0.98	0.99
disgust	0.99	1.00	0.99
fear	0.99	1.00	1.00
happy	0.99	0.99	0.99
neutral	1.00	1.00	1.00
sad	1.00	1.00	1.00
accuracy			0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99

EMODB

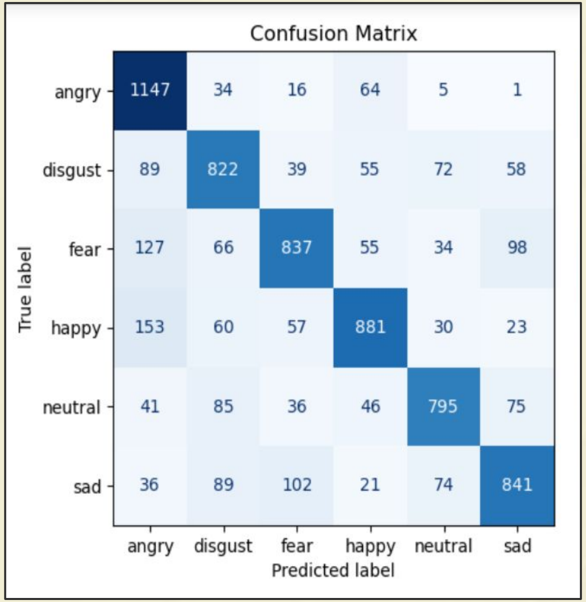
Accuracy (Boosted SVM and RF): 0.878504672897

	precision	recall	f1-score
angry	0.89	0.95	0.92
boredom	0.89	0.92	0.90
disgust	0.86	0.76	0.81
fear	0.76	0.97	0.85
happy	0.89	0.73	0.80
neutral	0.89	0.88	0.88
sad	0.96	0.81	0.88
accuracy			0.88
macro avg	0.88	0.86	0.87
weighted avg	0.88	0.88	0.88

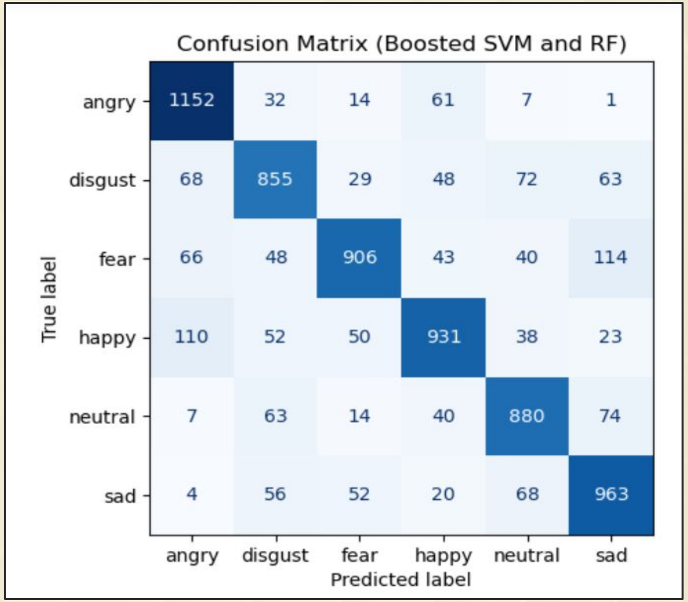
Results on Combined Dataset



Random Forest



SVM



Random Forest+SVM

Insights

Traditional Models with Proper Tuning:

- **Hyperparameter Optimization**: Enhanced performance by fine-tuning model parameters.
- **Ensemble Learning**: Leveraged multiple models to reduce individual weaknesses and improve prediction accuracy.

Performance:

- **Accuracy**: Comparable to advanced deep learning models (e.g., CNN, RNN).
- **Complexity**: Simplified model architecture while maintaining high performance.

Robustness:

- **Speaker Diversity**: Age, gender, accents, languages, recording environments, noise levels, audio quality.
- **Generic Applicability**: Effective across combined and individual datasets for Speech Emotion Recognition (SER).

Accuracy (Boosted SVM and RF): 0.8050679501698754

	precision	recall	f1-score	support
angry	0.82	0.91	0.86	1267
disgust	0.77	0.75	0.76	1135
fear	0.85	0.74	0.79	1217
happy	0.81	0.77	0.79	1204
neutral	0.80	0.82	0.81	1078
sad	0.78	0.83	0.80	1163
accuracy			0.81	7064
macro avg	0.81	0.80	0.80	7064
weighted avg	0.81	0.81	0.80	7064

Unlike in previous research papers, here, all emotions have good individual f1-score. Limiting bias.

06

FUTURE WORKS

Future Works

01

Add more Data Augmentation techniques such as Shifting and Pitch change

02

Comparison with Deep Learning Models

Classification report for Emotion Recognition

	precision	recall	f1-score
0	0.82	0.85	0.84
1	0.75	0.69	0.72
2	0.77	0.73	0.75
3	0.75	0.73	0.74
4	0.74	0.81	0.77
5	0.72	0.74	0.73
accuracy			0.76
macro avg	0.76	0.76	0.76
weighted avg	0.76	0.76	0.76

03

Nuanced understanding of feelings based on probabilistic emotion expression

Threats of Validity

Internal	External
• Selection Bias • Overfitting • Labelling Errors • Data Augmentation	• Might not reflect real world scenarios • Confounding Variables (quality of audio, speaker health) • Interaction Effects

Conclusion

- ➢ Reviewed 12 papers and focused on the various models and datasets used
- ➢ Found research gaps as the model was built on single dataset leading to the issue of lack of generalizability.
- ➢ The approach to improve this was by multi-dataset and combination of simple classification models rather than deep learning models
- ➢ Merged 5 datasets, 12k entries and 6 emotions
- ➢ Through Ensemble learning, combining the strengths of SVM and RF, achieved an accuracy of 81% and an 80% classification rate for each emotion
- ➢ In the future, we would want to explore more on hyperparameter tuning and introducing more languages and emotions

Thank You!!
Any questions?