BANGLADESH UNIVERSITY OF
ENGINEERING AND TECHNOLOGY

SPEECH EMOTION RECOGNITION

CSE 472 - MACHINE LEARNING SESSIONAL

Group - **A1_8**
Md Nabil Sadique - 1905006
Md. Huzzatun Ali - 1905027

PROBLEM DEFINITION

Machines often lack the ability to interpret emotional cues, limiting their effectiveness in human-centric applications. Speech Emotion Recognition can alleviate this problem.

Speech Emotion Recognition (SER) focuses on **identifying human emotions from speech signals.**

USE CASE OF SER

- Enhanced Human-Computer Interaction
- Healthcare Applications
- Customer Service
- Entertainment

DATASET

We'll use the **IEMOCAP dataset** which is a **12-hour collection of audio-visual recordings** of actors in dialogues, **labeled with emotions** (e.g., happiness, anger) **as the base dataset**. It's widely used for developing and testing speech emotion recognition models.

From this base dateset, we chose the pre-processed variant which has the **audio data**, **spectrogram**, **MFCC (Mel-frequency cepstral coefficient).**

DATASET

IEMOCAP Full-release

https://www.kaggle.com/datasets/dejolilandry/iemocapfullrelease

IEMOCAP Pre-processed Dataset

https://drive.google.com/file/d/1Nnxh3y7hkkmsh3Y5Dg4q1qerRZWcePH8/view

DATASET ANALYSIS

Why IEMOCAP

Database	Speech Emotion Category						
	Surprise	Angry	Happy	Fear	Sad	Neutral	Disgust
EmoDB	—	127	71	69	62	79	81
eNTERFACE	215	215	212	215	215	—	215
AFEW4.0	103	156	171	113	145	167	106
IEMOCAP	—	1103	1636	—	1084	1708	—

DATASET ANALYSIS

IEMOCAP's emotion distribution

Emotion	Number of Examples
Angry	1103
Happy+Excited	1636
Neutral	1708
Sad	1084
Total	**5531**

EXISTING METHOD

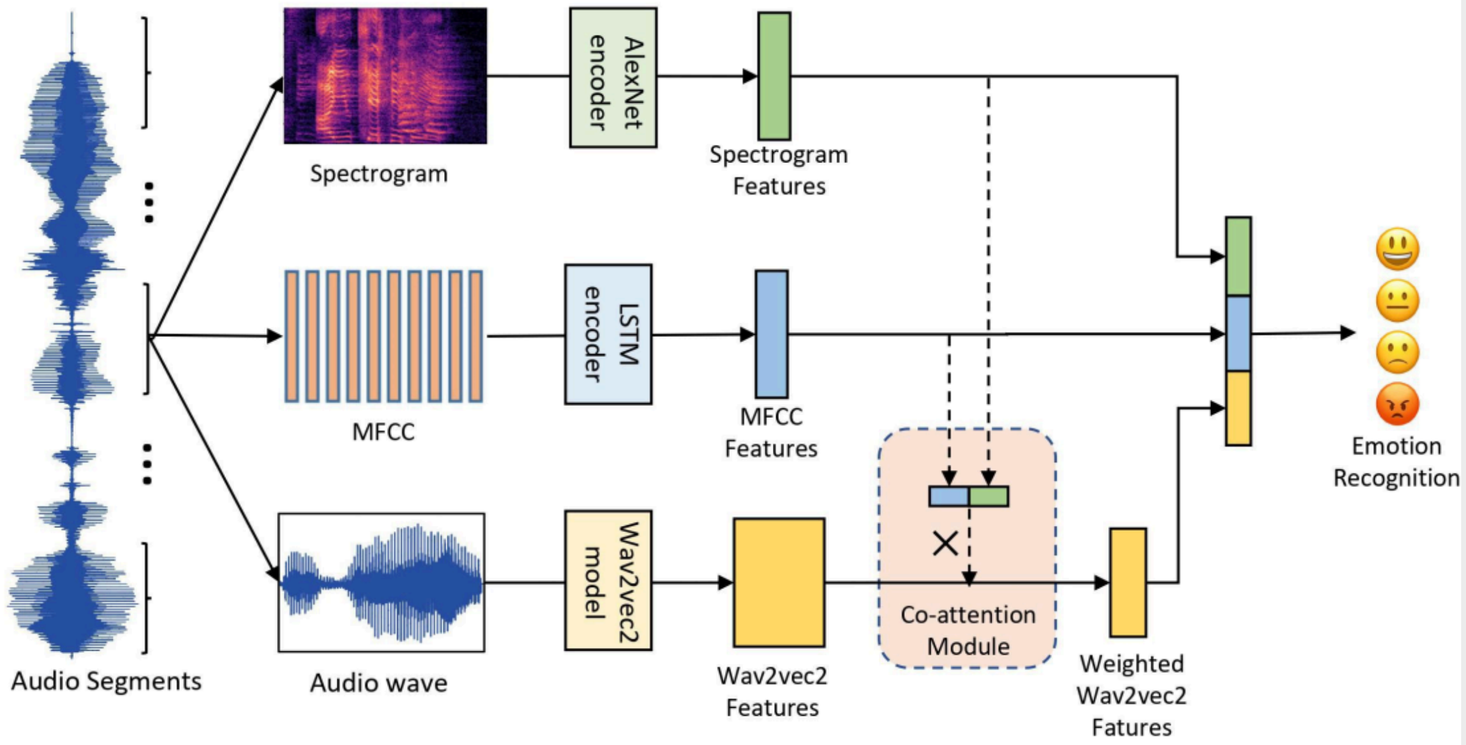


Fig. 1. The overall architecture of our proposed method.

SPEECH EMOTION RECOGNITION WITH CO-ATTENTION BASED MULTI-LEVEL ACOUSTIC INFORMATION

PROPOSED SOLUTION

Our Proposal -

- We have incorporated **Vision Transformer instead of AlexNet Encoder.**
- We have done **ensembling by majority voting** of the output of different models.

PROPOSED SOLUTION

Step -1 (Vision Transformer instead of AlexNet Encoder)

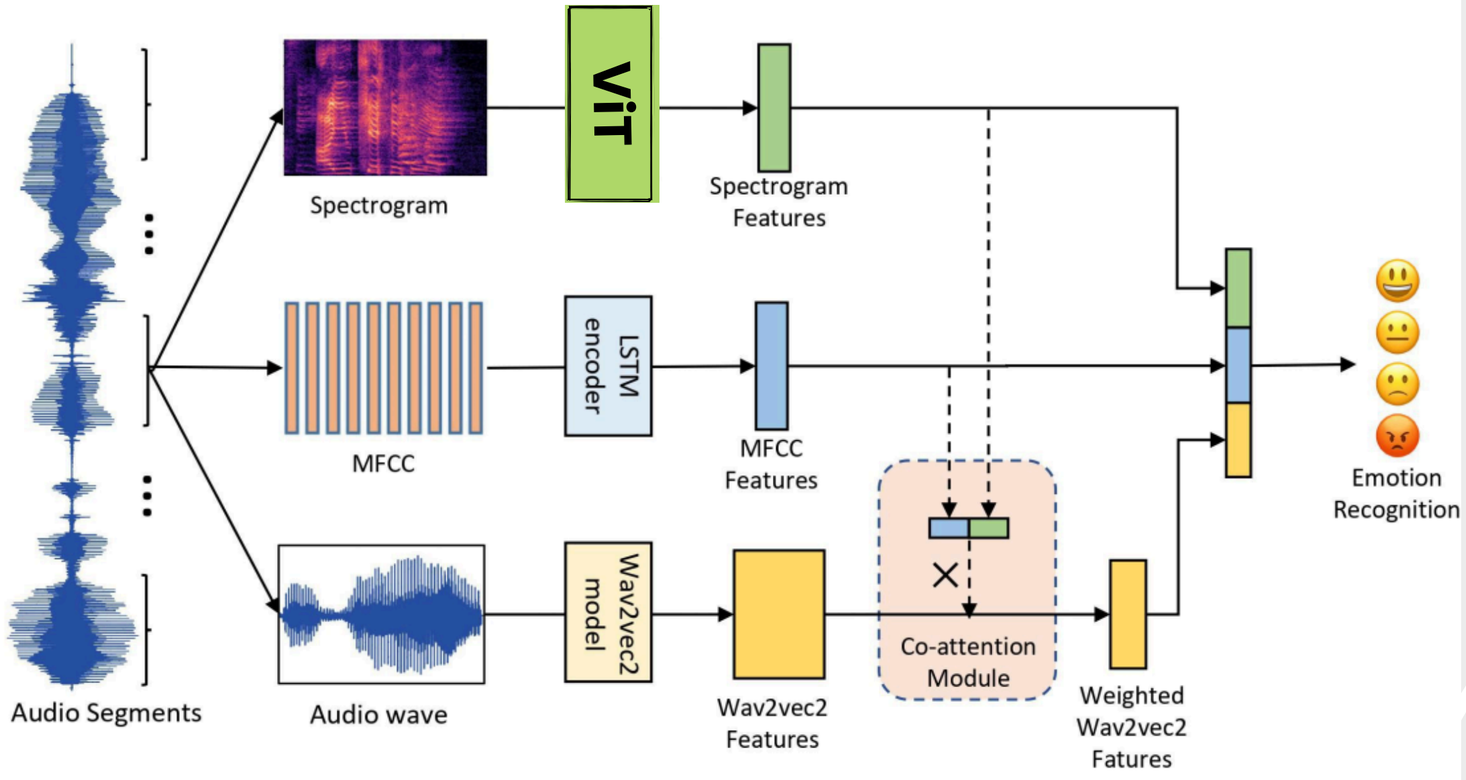


Fig. 1. The overall architecture of our proposed method.

PROPOSED SOLUTION

Step -2 (Ensembling with Majority Voting)

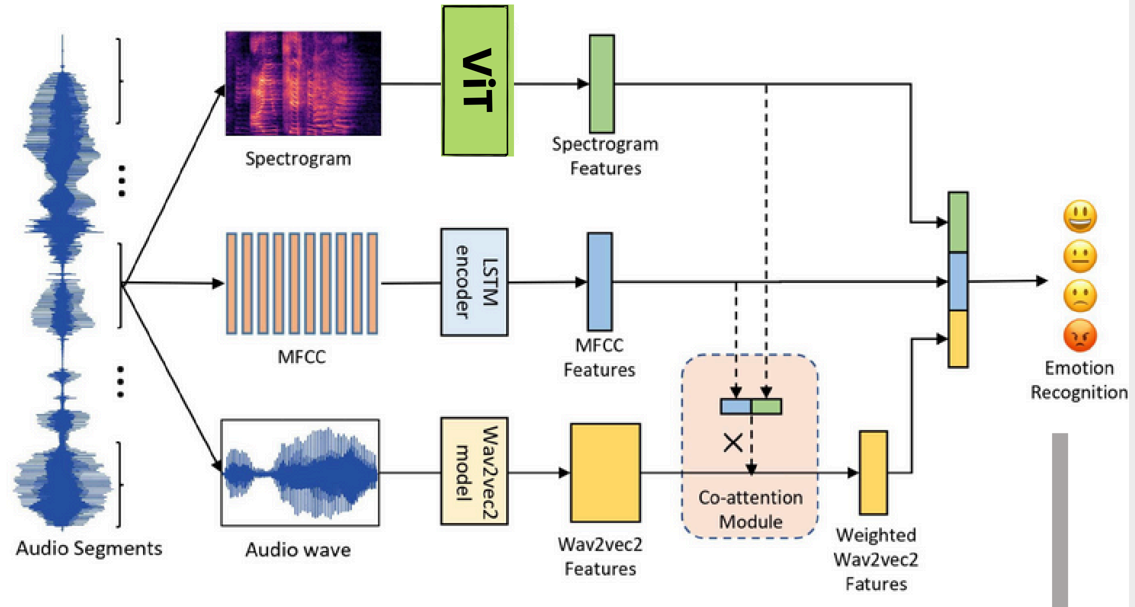


Fig. 1. The overall architecture of our proposed method.

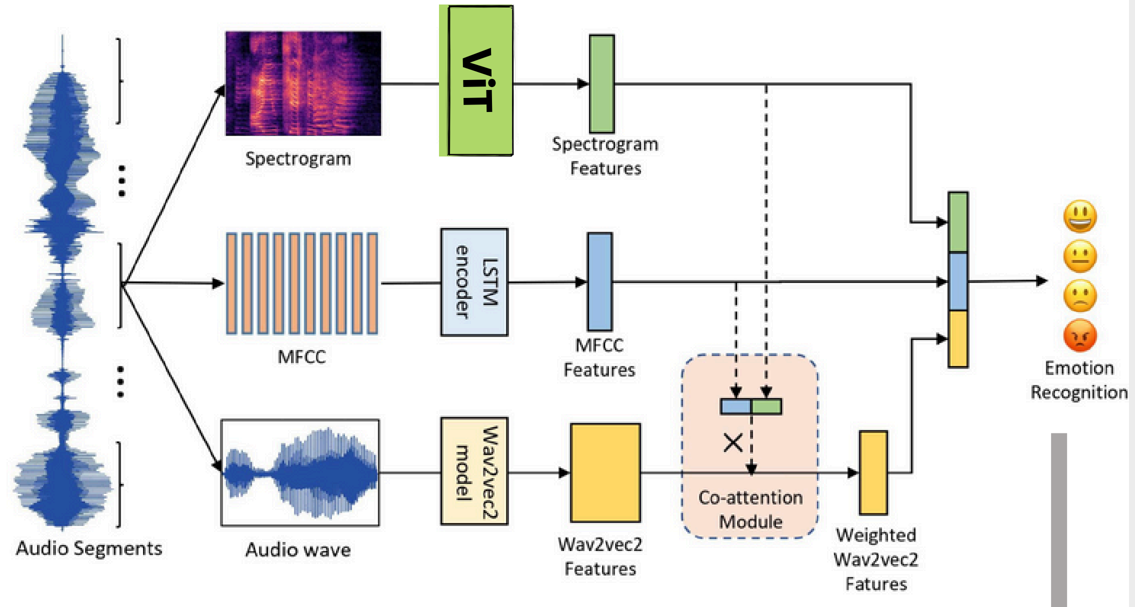


Fig. 1. The overall architecture of our proposed method.

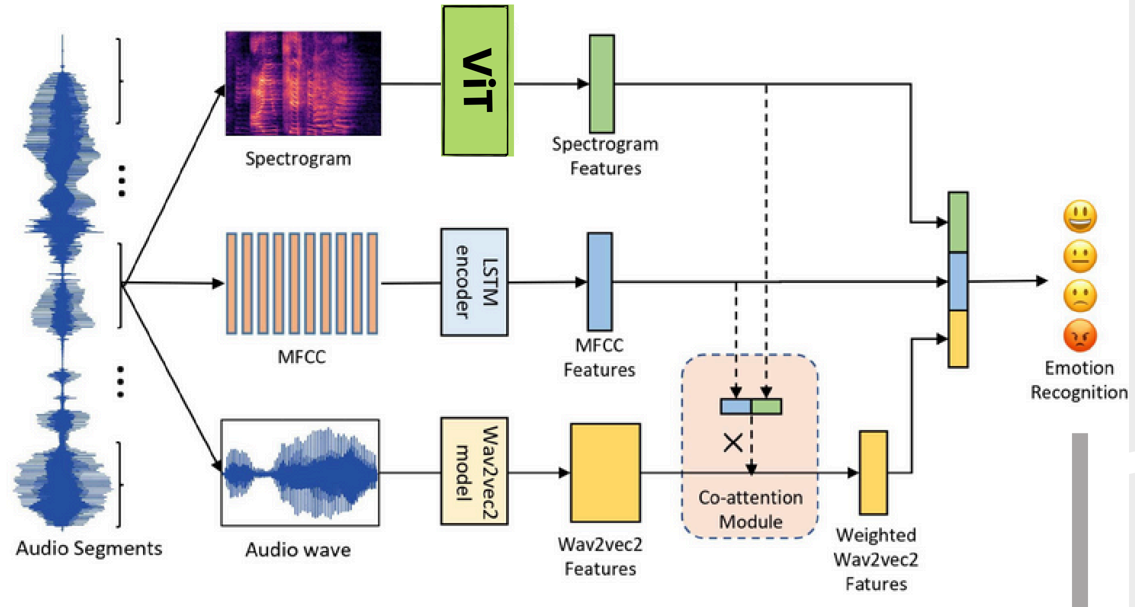
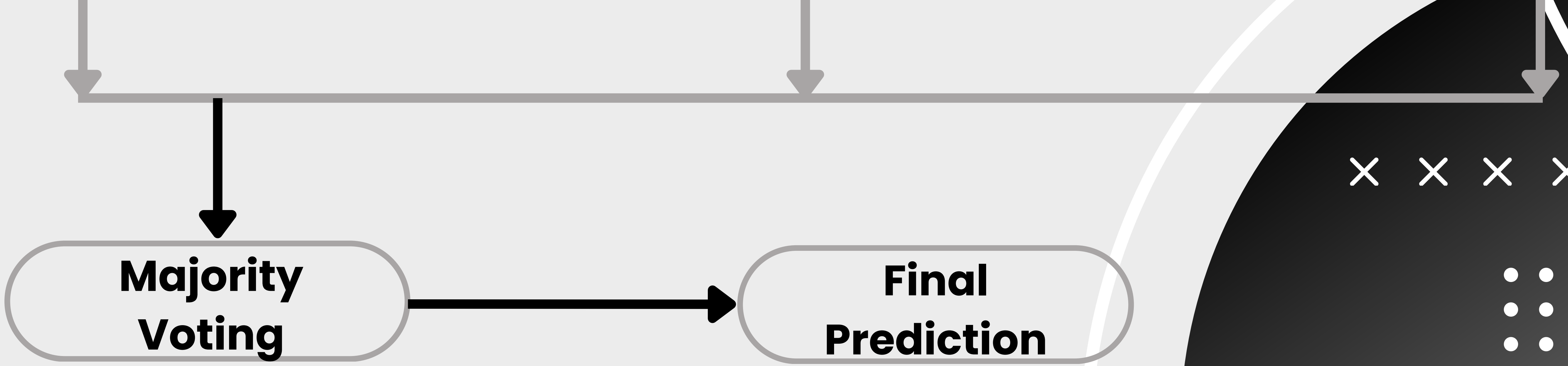


Fig. 1. The overall architecture of our proposed method.



Majority Voting

Final Prediction

LOSS FUNCTION <<<<<

1. **Cross Entropy Loss -**

As it is a multiclass classification, the loss function used here is cross entropy loss

>>>>>

PERFORMANCE METRICS

1 Unweighted Accuracy -

We can measure it by the **total loss divided by total number of samples**

As it is multiclass classification problem we can get the avg loss by UA loss. **It gives same importance to each class.**

PERFORMANCE METRICS

2 Weighted Accuracy -

It gives more importance to the class with fewer samples to **maintain the coherence in emphasizing each class**.

In case of **imbalanced dataset**, we need this loss function to **ensure the importance of each class properly**

PERFORMANCE REPORT

- 7 different models have been trained by selecting 5 actors from the first 9 actors of the dataset randomly.
- 4 of the actors are taken as training set and the reamining is taken as validation set.
- The last(10th) actor is considered as the test data.
- We ran the models on the test data and did majority voting.

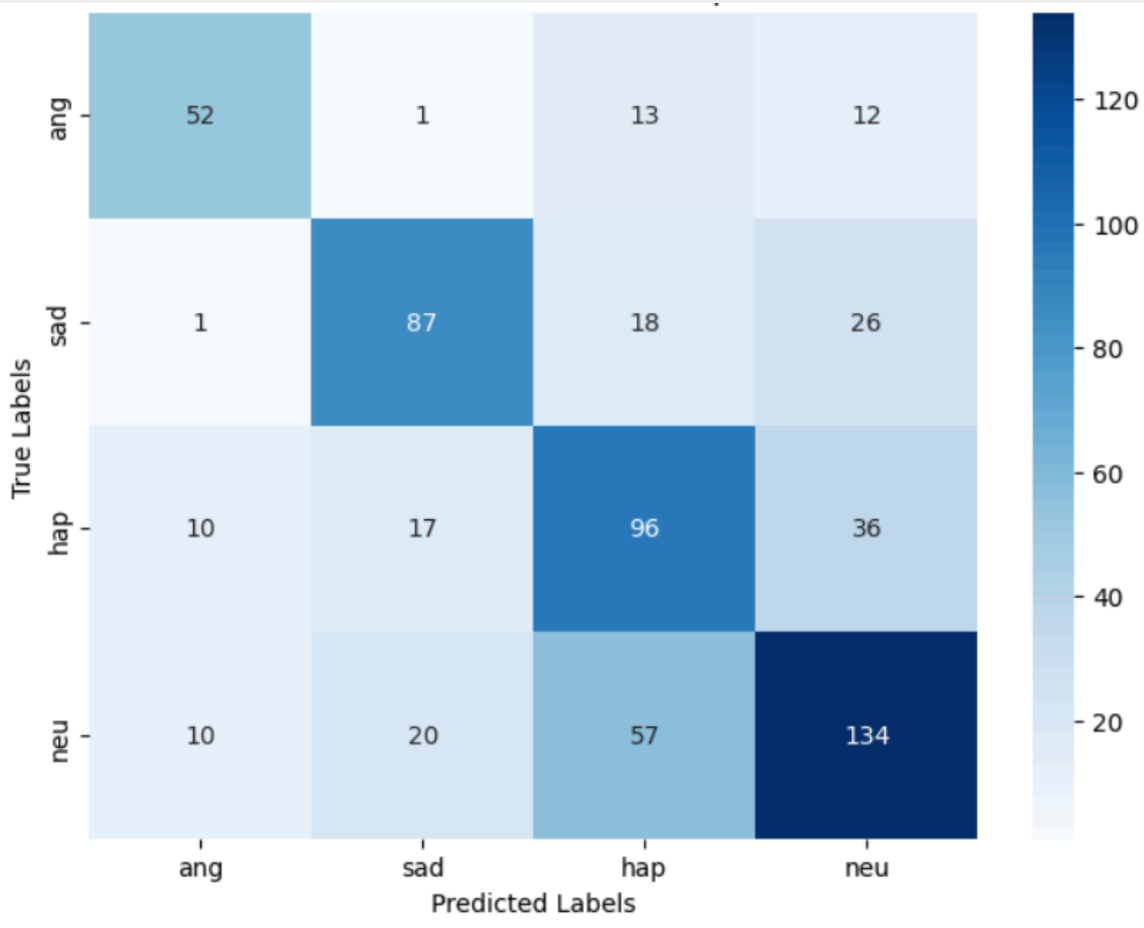
PERFORMANCE REPORT

RESULTS ON TEST SET AFTER ENSEMBLING FOR VIT:

```
WA: 62.54
```

```
UA: 63.40
```

PERFORMANCE REPORT



COMPARISON

RESULTS ON TEST SET AFTER ENSEMBLING FOR VIT:

WA: 62.54

UA: 63.40

<

>

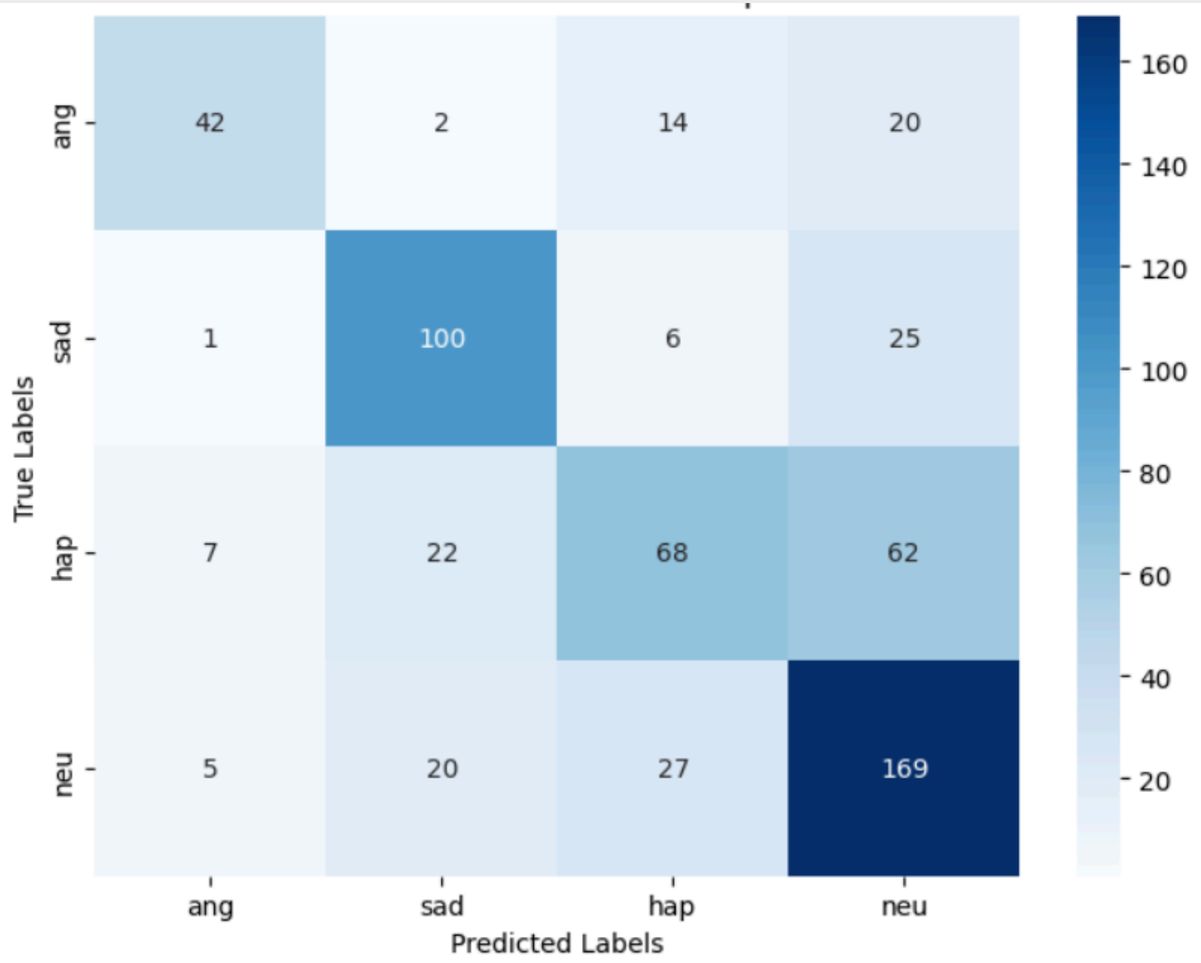
RESULTS ON TEST SET AFTER ENSEMBLING FOR CNN:

WA: 64.24

UA: 62.21

COMPARISON

AFTER ENSEMBLING FOR CNN



COMPARISON

RESULTS ON TEST SET AFTER ENSEMBLING FOR VIT:

WA: 62.54

UA: 63.40

<

>

RESULTS ON TEST SET (WITHOUT ENSEMBLING) FOR CNN:

WA: 64.24

UA: 62.21

DISCUSSION

RESULTS ON TEST SET (WITHOUT ENSEMBLING) FOR CNN:

MODEL-1

WA: 64.07 UA: 65.75

MODEL-2

WA: 64.75 UA: 62.05

MODEL-3

WA: 64.24 UA: 61.48

MODEL-4

WA: 64.24 UA: 62.16

MODEL-5

WA: 63.39 UA: 60.95

MODEL-6

WA: 61.53 UA: 60.34

MODEL-7

WA: 63.05 UA: 66.97

DISCUSSION

The spectogram is of **384*256 shape**. But due to time and resource constraint we **interpolate it to a 224*224 shape**. And then we ran our proposed model on it.

Due to **reshaping the image some information has been lost**. Still our model performs nearly similar to the state-of-the-art model.

DISCUSSION

In the paper, they have **splitted the dataset into 8:1:1** as **the train, validation and test dataset**

But due to resource constraint, **we split the dataset into 4:1:1 as the train, validation and test set**

DISCUSSION

The difference in result in different training cases shows that there is **a personal bias** while creating a dataset for Speech Emotion Detection.

So, without ensembling, **taking the direct result** from 1-2 or the best performing model might result **in huge success in similar cases and huge degradation of performance in opposing cases.** So, **there will be inconsistency without ensembling.**

THANK YOU