SPEECH EMOTION RECOGNITION

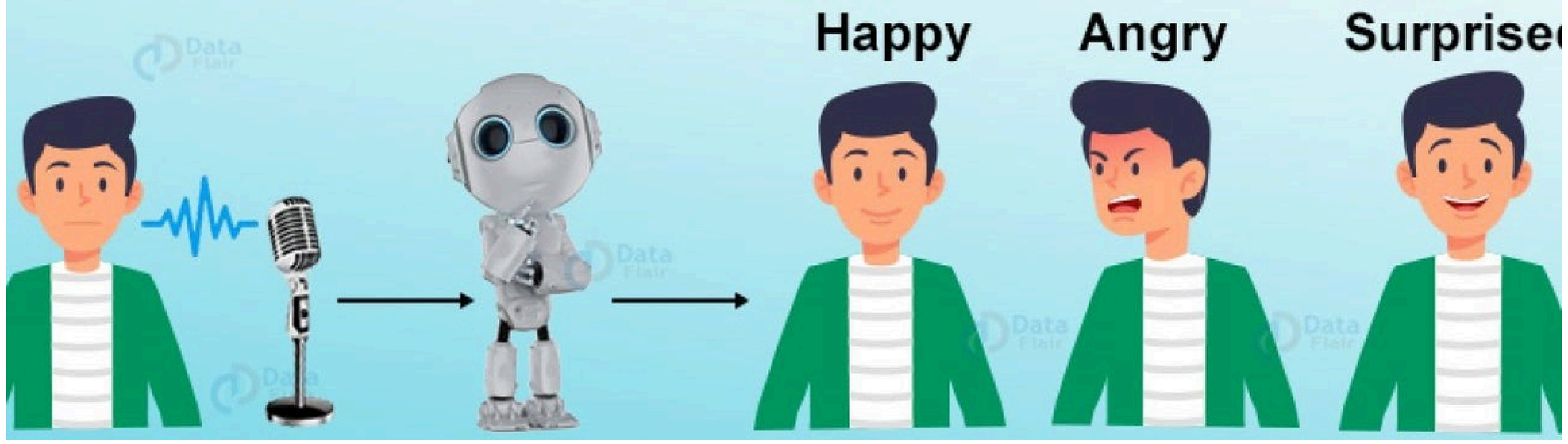
SPEECH & EMOTIONS

Speech is the natural and widespread human communication method and carries paralinguistic and linguistic information.

Linguistic information includes the context and language of the speech, while paralinguistic information includes the gender, emotions, age, and other unique attributes of the human.

Emotion is the cornerstone of human interactions. It is our brain's creation of what our bodily sensations mean, in relation to what is going on around you in the world

WHAT'S SPEECH EMOTION RECOGNITION ?



Speech Emotion Recognition(SER) is a task of speech processing and computational paralinguistics that aims to recognize and categorize the emotions expressed in spoken language.

The goal is to determine the emotional state of a speaker, such as happiness, anger, sadness, or frustration, from their speech patterns, such as prosody, pitch, and rhythm.

DATASET - RAVDESS

Ryerson Audio-Visual Database of Emotional Speech and Song dataset consists of speech audio-only files.

It contains 60 trials per actor x 24 actors = 1440.
The voices contain 24 professional actors (12 female, 12 male), vocalizing two lexically matched statements in a neutral North American accent.

Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

DATA PREPARATION

We collected data from the RAVDESS dataset, which includes labeled audio recordings representing various emotions.

This involves listing directories, parsing file names to extract emotion labels, and creating DataFrames for emotions and file paths, ultimately resulting in a comprehensive dataset for our deep learning model.

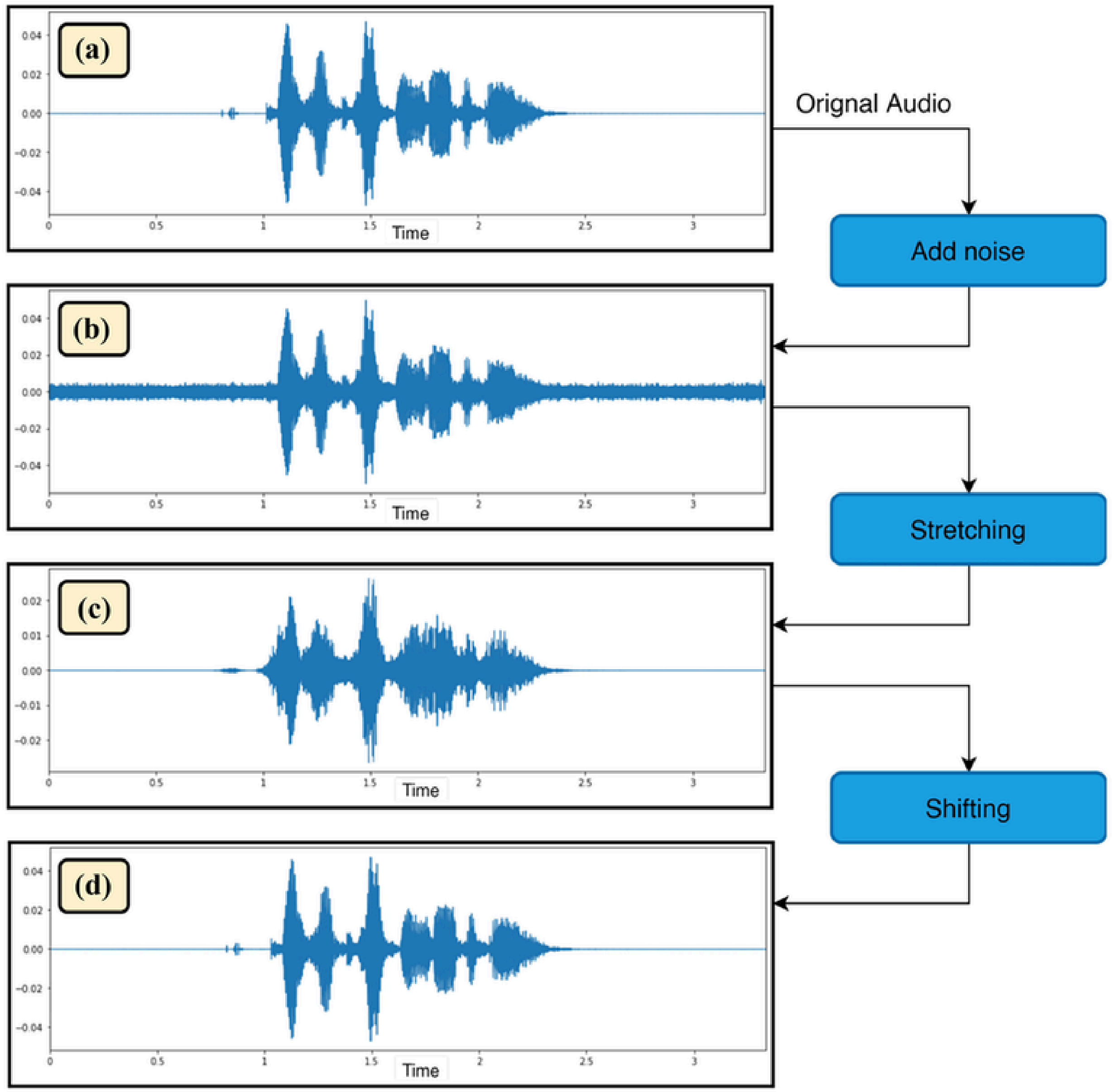
DATA AUGMENTATION

To enhance the dataset for our project, we applied two data augmentation techniques: adding noise and shifting data.

The noise function introduces random noise based on the maximum amplitude of the audio, replicating real-world variations.

The shift function randomly changes the starting point of the audio, simulating time-based variations.

These augmentations increase dataset diversity, leading to a more robust and adaptable deep learning model, while reducing the risk of overfitting.



DATA PREPROCESSING

To prepare data for our speech emotion recognition project, we first created a data frame of features and one-hot encoded labels, then split the data into training and testing sets for model evaluation.

The data was standardized using 'StandardScaler' to ensure consistency and then reshaped to meet the input requirements of our deep learning model.

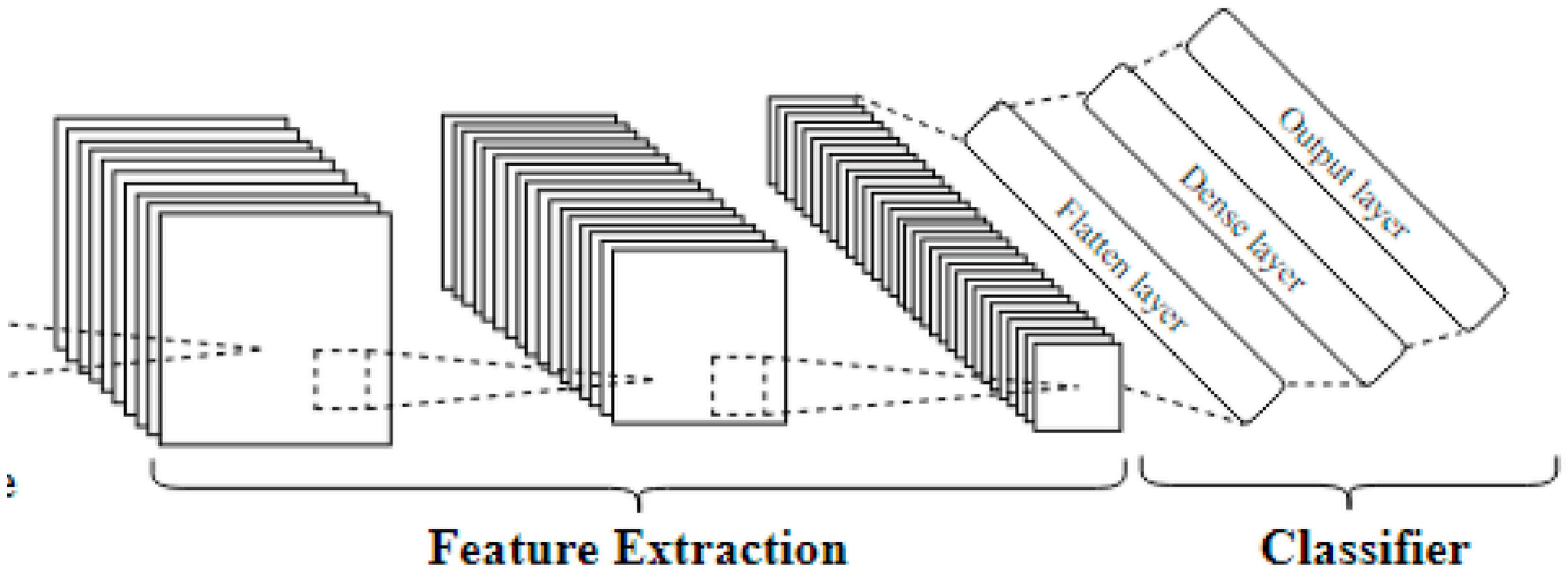
These steps ensure our data is properly structured, scaled, and formatted for training and testing.

MODEL ARCHITECTURE

We designed a deep learning model for speech emotion recognition using Convolutional Neural Networks (CNNs).

The model consists of four Conv1D layers, each followed by MaxPooling1D for downsampling and Dropout for regularization. After flattening the feature maps, the Dense layers include a hidden layer with 32 units and a final output layer with 8 units, corresponding to the eight possible emotions.

The model is compiled with the Adam optimizer and categorical crossentropy loss.



TRAINING & TESTING

To optimize our model training, we used two key callbacks: ReduceLROnPlateau, which reduces the learning rate when the monitored loss metric stops improving, and Model Checkpoint, which saves the best model based on validation loss.

We trained our SER model for 60 epochs with a batch size of 64, using validation data to monitor progress. After training, we evaluated the model's accuracy on the test set, achieving a result of 85.55 %.

EVALUATION METRICS

The model is evaluated on the 4 different evaluation metrics Accuracy, precision, recall and flscore. The confusion matrix is used to calculate the evaluation parameters from which the metrics have been calculated.

ACCURACY

TP + TN /
(TP + TN + FP + FN)

RECALL

TP /
TP + FN

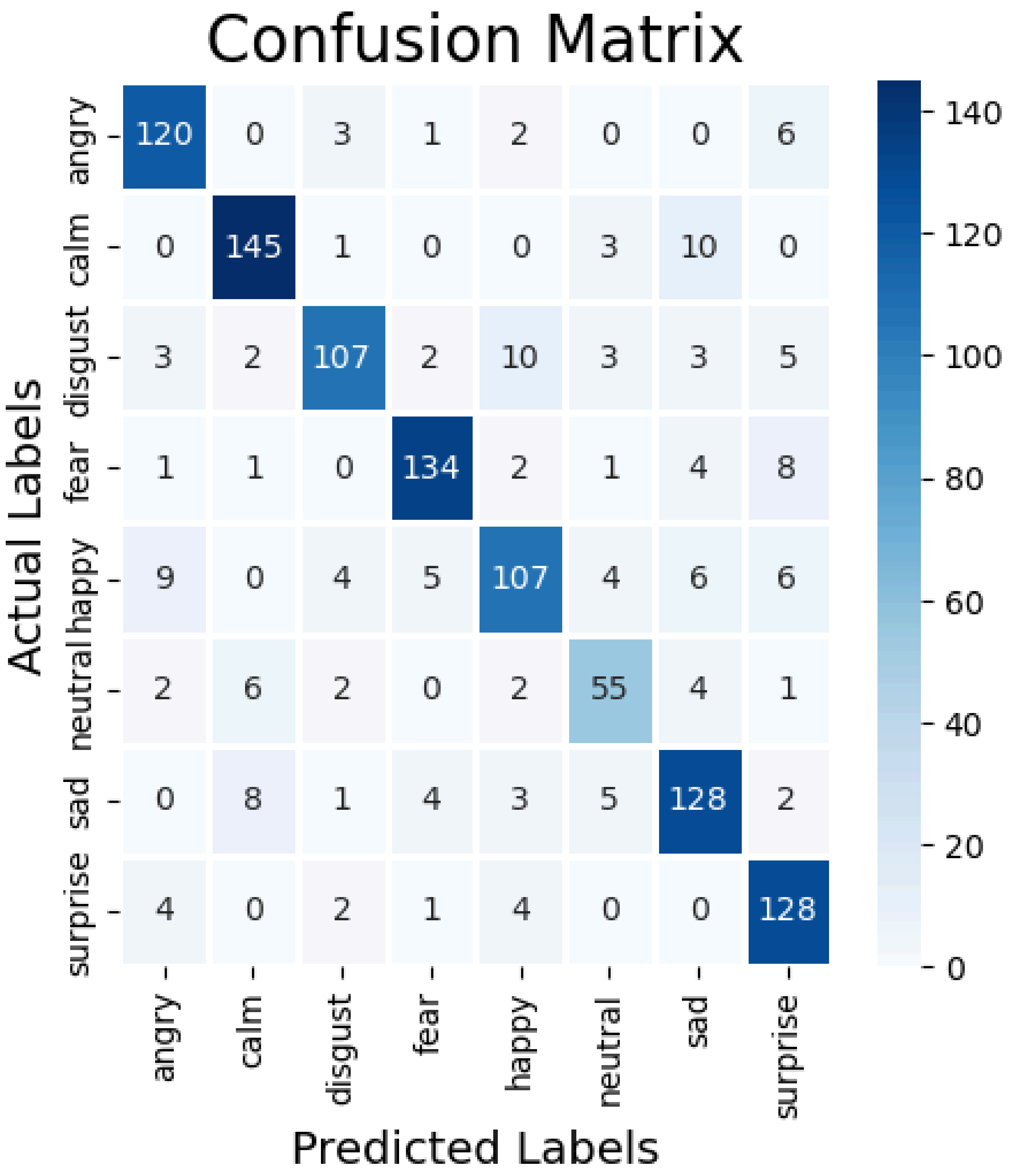
PRECISION

TP /
TP + FP

F1 SCORE

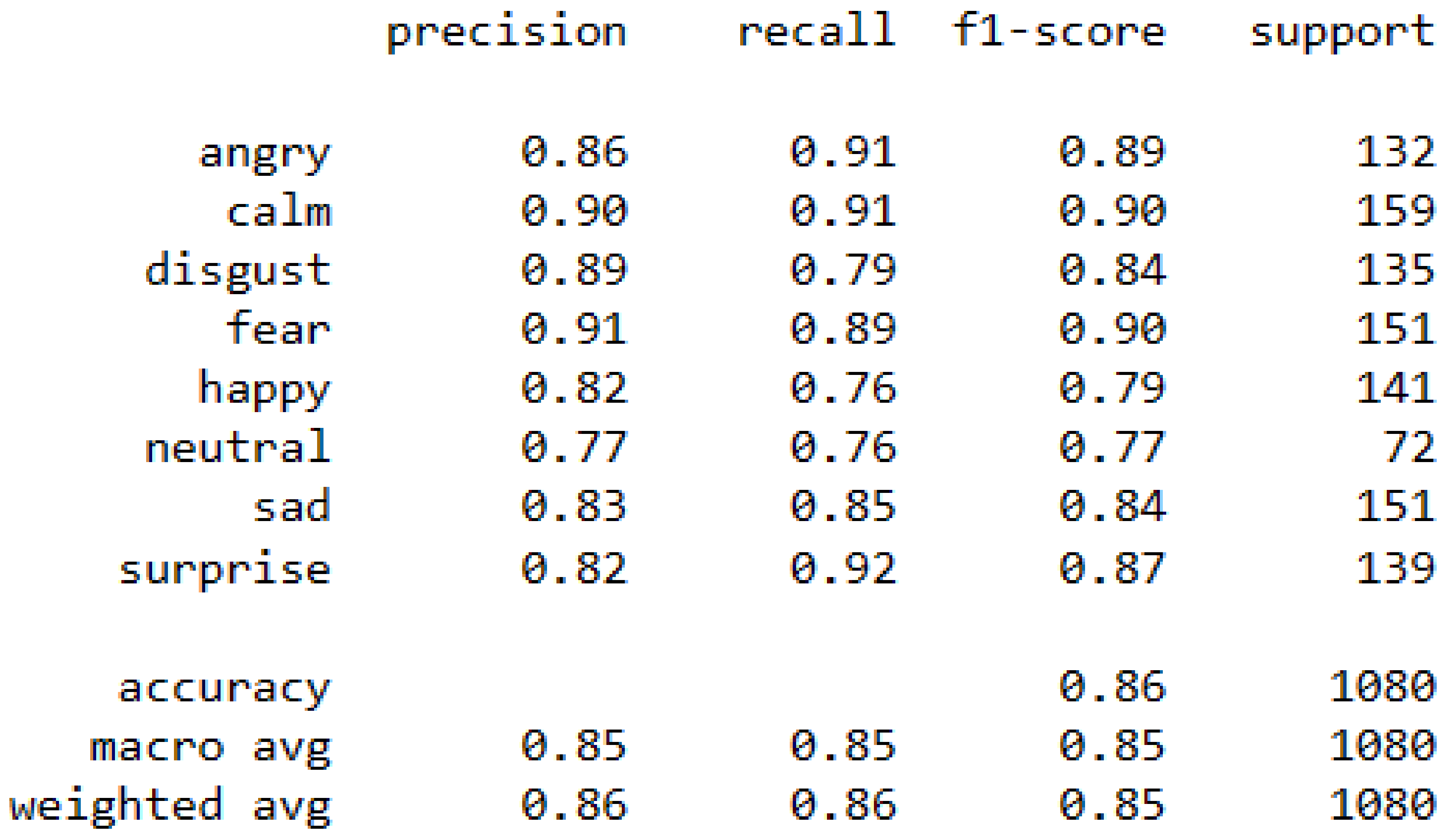
2 X [(PRECISION X RECALL)/
(PRECISION + RECALL)]

RESULTS



Confusion Matrix

	angry	calm	disgust	fear	happy	neutral	sad	surprise
angry	120	0	3	1	2	0	0	6
calm	0	145	1	0	0	3	10	0
disgust	3	2	107	2	10	3	3	5
fear	1	1	0	134	2	1	4	8
happy	9	0	4	5	107	4	6	6
neutral	2	6	2	0	2	55	4	1
sad	0	8	1	4	3	5	128	2
surprise	4	0	2	1	4	0	0	128

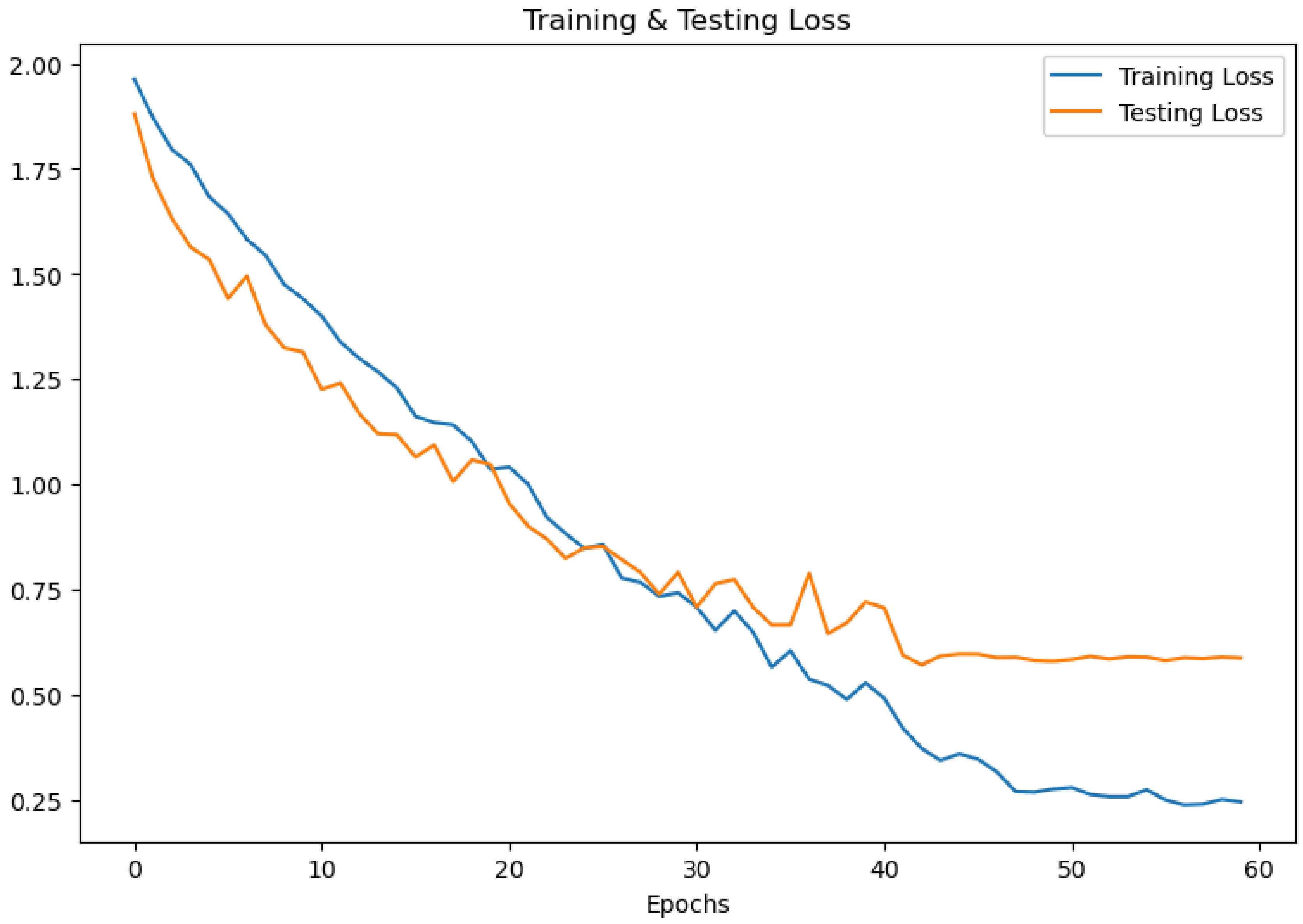


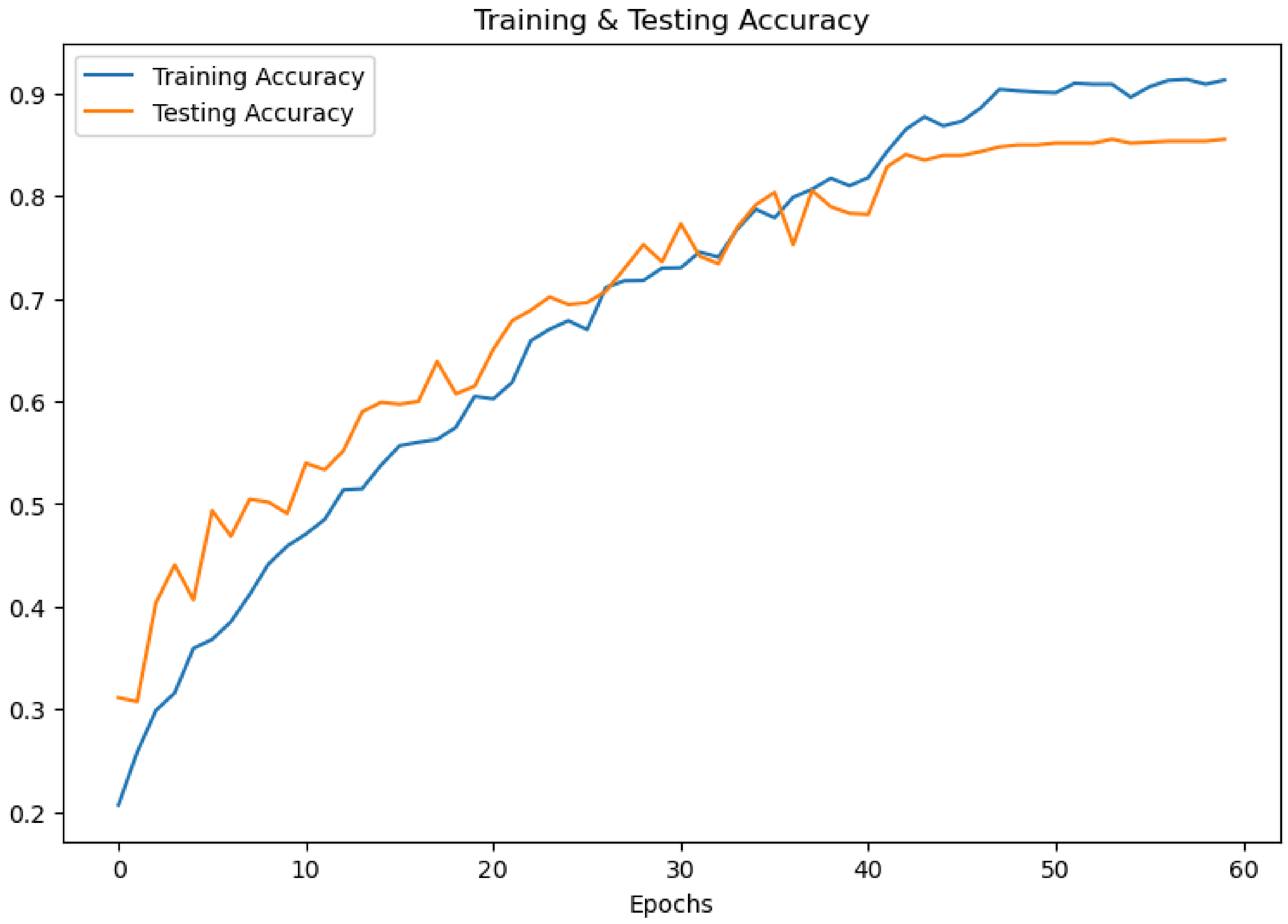
```
              precision    recall  f1-score   support

       angry       0.86      0.91      0.89       132
        calm       0.90      0.91      0.90       159
     disgust       0.89      0.79      0.84       135
        fear       0.91      0.89      0.90       151
       happy       0.82      0.76      0.79       141
     neutral       0.77      0.76      0.77        72
         sad       0.83      0.85      0.84       151
    surprise       0.82      0.92      0.87       139

    accuracy                           0.86      1080
   macro avg       0.85      0.85      0.85      1080
weighted avg       0.86      0.86      0.85      1080
```

ACCURACY & LOSS PLOTS

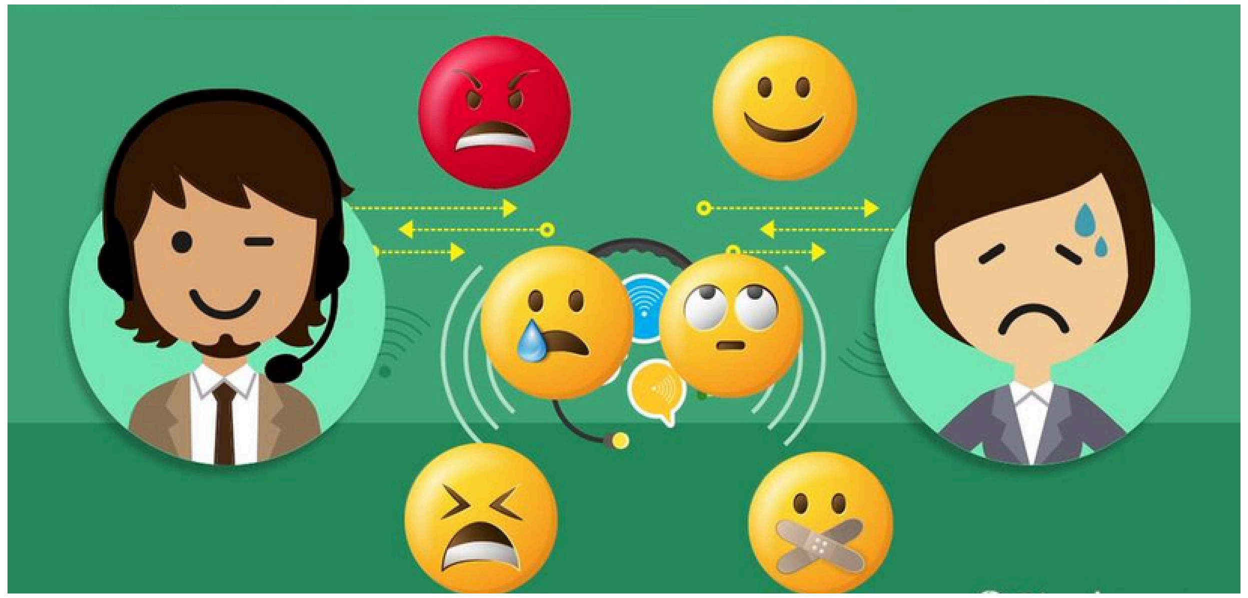

Training & Testing Loss

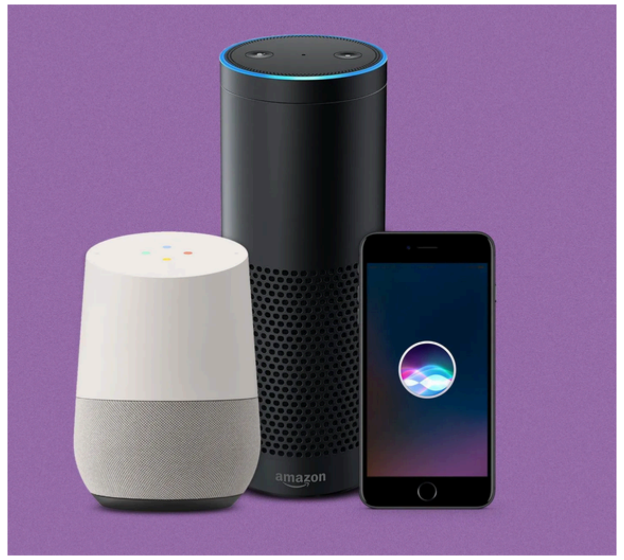

Training & Testing Accuracy

APPLICATIONS OF SER

CUSTOMER SERVICE

In customer service, SER can be used to detect customer emotions during phone calls or chat interactions. This helps customer service representatives adapt their responses to the caller's emotional state, resulting in better customer satisfaction.





VIRTUAL ASSISTANTS

Virtual assistants like Siri, Alexa, or Google Assistant can use SER to understand user emotions. This enables them to respond more empathetically and create a more personalized experience.

APPLICATIONS OF SER

MENTAL HEALTH MONITORING

SER can assist mental health professionals in monitoring emotional changes in patients. It can detect shifts in mood, stress levels, or other emotional indicators that may signal the need for intervention.

CAR SAFETY SYSTEMS

SER algorithms can analyze the driver's speech patterns, tone, and voice characteristics in real-time to determine their emotional state while driving.
If signs of distraction, drowsiness, stress, or other negative emotions are detected, the system can generate alerts to notify the driver and potentially prevent accidents.

THANK YOU