Data Science Tool Workshop
CDCSC19

Project 1 (TinyML)
Emotion Echo – A Speech Emotion Recognition Model

Team -

Saharsh (2022UCD2109)

Ritesh Wadhwani (2022UCD2149)

Aakash Arora (2022UCD2118)

Overview-

The speech emotion on recognition model utilizes audio feature extraction techniques like **MFCC, Chroma-stft, ZCR, RMS, and Mel-spectrogram**, combined with data augmentation methods such as noise addition, pitch shifting, and stretching, to enhance data diversity. It employs a **Convolutional Neural Network (CNN)** architecture with multiple Conv1D layers, Batch Normalization, Dropout, MaxPooling, and a Dense softmax layer for multiclass classification across seven emotions. The model was trained using the **Adam optimizer** and categorical cross-entropy loss, achieving strong performance with effective training-validation accuracy and loss convergence. Evaluated on a test set, it demonstrated reliable emotion classification, supported by robust metrics and a detailed classification report, with opportunities for further improvement through advanced architectures, hyperparameter tuning, and additional data.

This model is then converted **.tflite** file in order to demonstrate **TinyML** on Edge Impulse.

1. **Dataset and Feature Extraction:**
   - **Dataset:** We used an audio dataset containing speech data with associated emo on labels.
   - **Feature Extraction:** The features extracted from the audio include:
     - **Zero Crossing Rate (ZCR):** A measure of the rate at which the signal changes its sign.
     - **Chroma-stft:** A representation of the energy distribution in pitch classes over me.
     - **MFCC (Mel Frequency Cepstral Coefficients):** Captures the timbral texture of the audio signal.
     - **Root Mean Square (RMS):** Represents the energy of the audio signal.
     - **Mel-Spectrogram:** Represents the spectral energy distribution over time.
   - **Augmentations:**
       - **Noise:** Random noise was added to the audio data.
       - **Stretching:** Time-stretching was applied to the audio to change the speed.
       - **Shifting:** The audio was shifted in time.
       - **Pitch Shift :** The pitch of the audio was shifted.
2. **Model Architecture:**
   - **Input Layer:** The input data is 1D time-series data (features extracted from the audio), reshaped into (samples, features, 1) for compatibility with the Conv1D layers.

Conv1D Layers: These layers capture spatial patterns in the 1D input:

- The network uses several **Conv1D layers** with 256, 128, and 64 filters of size 8.
- **Activation Function:** ReLU is used to introduce non-linearity.
- **Batch Normalization:** Helps in stabilizing the learning process.
- **Dropout:** Applied to reduce overfitting with a rate of 0.4.
- **MaxPooling1D:** Reduces the dimensionality of the data after each block of convolutional layers.
- **Flatten Layer:** After convolution, the output is flattened to be passed into the fully connected layers.
- **Dense Output Layer:** A so max layer with 7 units for multiclass classification (one for each emotion class). This is suitable for a multiclass classification problem where each input is assigned to one of the seven emotional classes.

3. **Model Compilation and Optimization:**

- **Optimizer:** Adam optimizer with a learning rate of 0.001.
- **Loss Function:** Categorical cross-entropy, which is suitable for multi class classification tasks.
- **Metrics:** The model uses accuracy (acc) as the evaluation metric.

4. **Model Training:**

- **Data Splitting:** The dataset was split into training and testing sets using a 75-25% ratio (using train_test_split).
- **Scaling:** The features were scaled using StandardScaler to ensure the data is on the same scale, improving the performance of the model.

- **Batch Size and Epochs:** The model was trained with a batch size of 64 for 100 epochs, with validation data used to monitor the performance during training.
- **Checkpoints:** Model checkpoints were used to save the best model based on the highest validation accuracy during training.

5. **Model Performance:**

- **Training History:** The model achieved good accuracy during training, with a notable improvement over epochs in both training and validation accuracy.
- **Loss Function:** The loss decreased steadily, indica ng the model was learning and improving during training.
- **Classification Report:**
    - The **precision, recall**, and **f1-score** for each class (emotion) show how well the model identifies emotions in the audio.
    - The **accuracy** score tells the overall performance of the model on the test set.
    - The **macro avg** and **weighted avg** provide the average performance across all classes, accounting for the imbalance of class distributions.

The ML Model-

The Google Colab ML code is attached below-

https://colab.research.google.com/drive/1hp5Se-gIZUlsiqCxNzbRatLAK6eU4y5a?usp=sharing

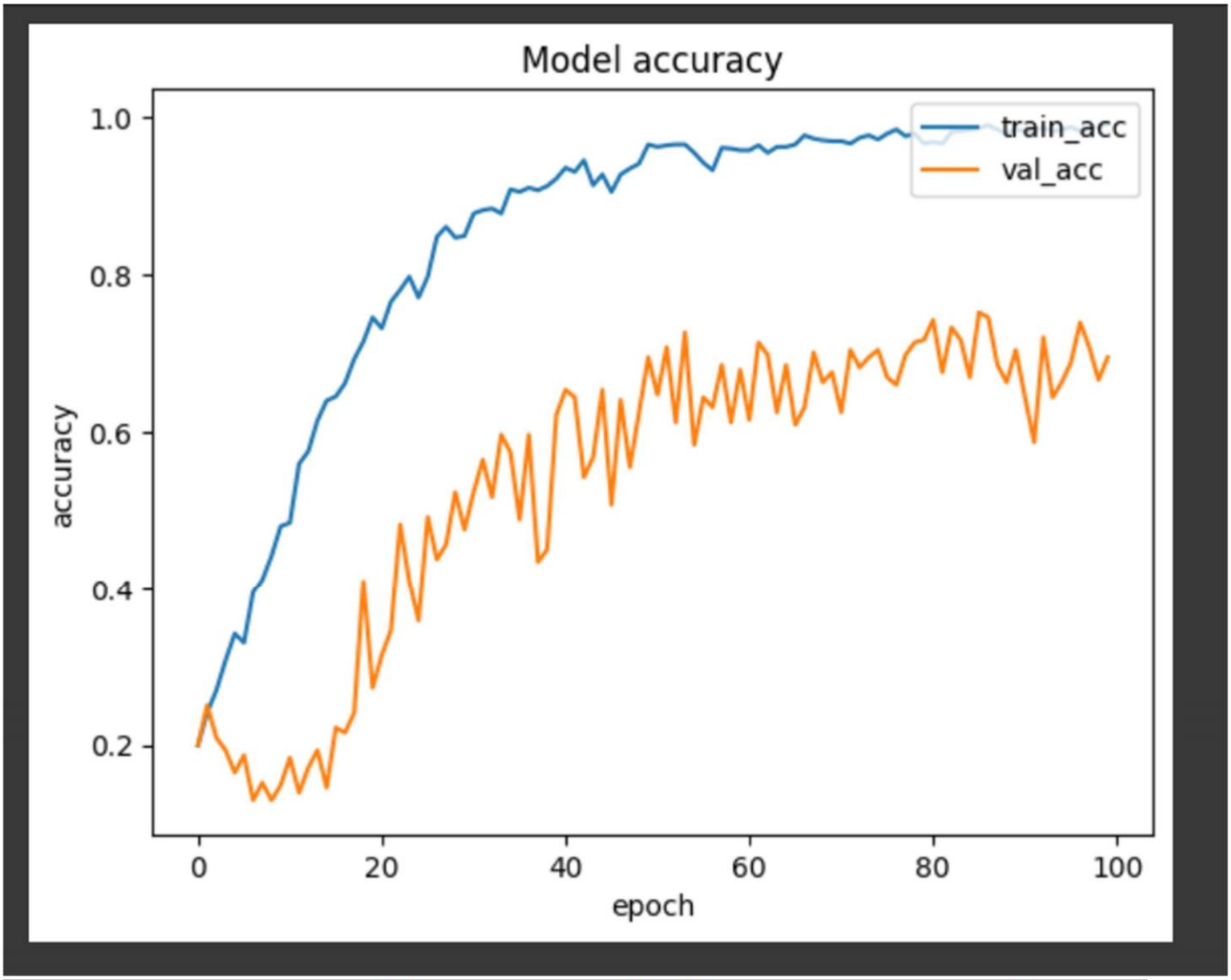
The CNN layers-

Model: "sequential_1"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 162, 256)	2,304
conv1d_1 (Conv1D)	(None, 162, 256)	524,544
batch_normalization (BatchNormalization)	(None, 162, 256)	1,024
dropout (Dropout)	(None, 162, 256)	0
max_pooling1d (MaxPooling1D)	(None, 20, 256)	0
conv1d_2 (Conv1D)	(None, 20, 128)	262,272
conv1d_3 (Conv1D)	(None, 20, 128)	131,200
dropout_1 (Dropout)	(None, 20, 128)	0
conv1d_4 (Conv1D)	(None, 20, 128)	131,200
conv1d_5 (Conv1D)	(None, 20, 128)	131,200
batch_normalization_1 (BatchNormalization)	(None, 20, 128)	512
dropout_2 (Dropout)	(None, 20, 128)	0
max_pooling1d_1 (MaxPooling1D)	(None, 2, 128)	0
conv1d_6 (Conv1D)	(None, 2, 64)	65,600
conv1d_7 (Conv1D)	(None, 2, 64)	32,832
conv1d_7 (Conv1D)	(None, 2, 64)	32,832
flatten_1 (Flatten)	(None, 128)	0
dense_1 (Dense)	(None, 7)	903

Total params: 3,849,239 (14.68 MB)
Trainable params: 1,282,823 (4.89 MB)
Non-trainable params: 768 (3.00 KB)
Optimizer params: 2,565,648 (9.79 MB)

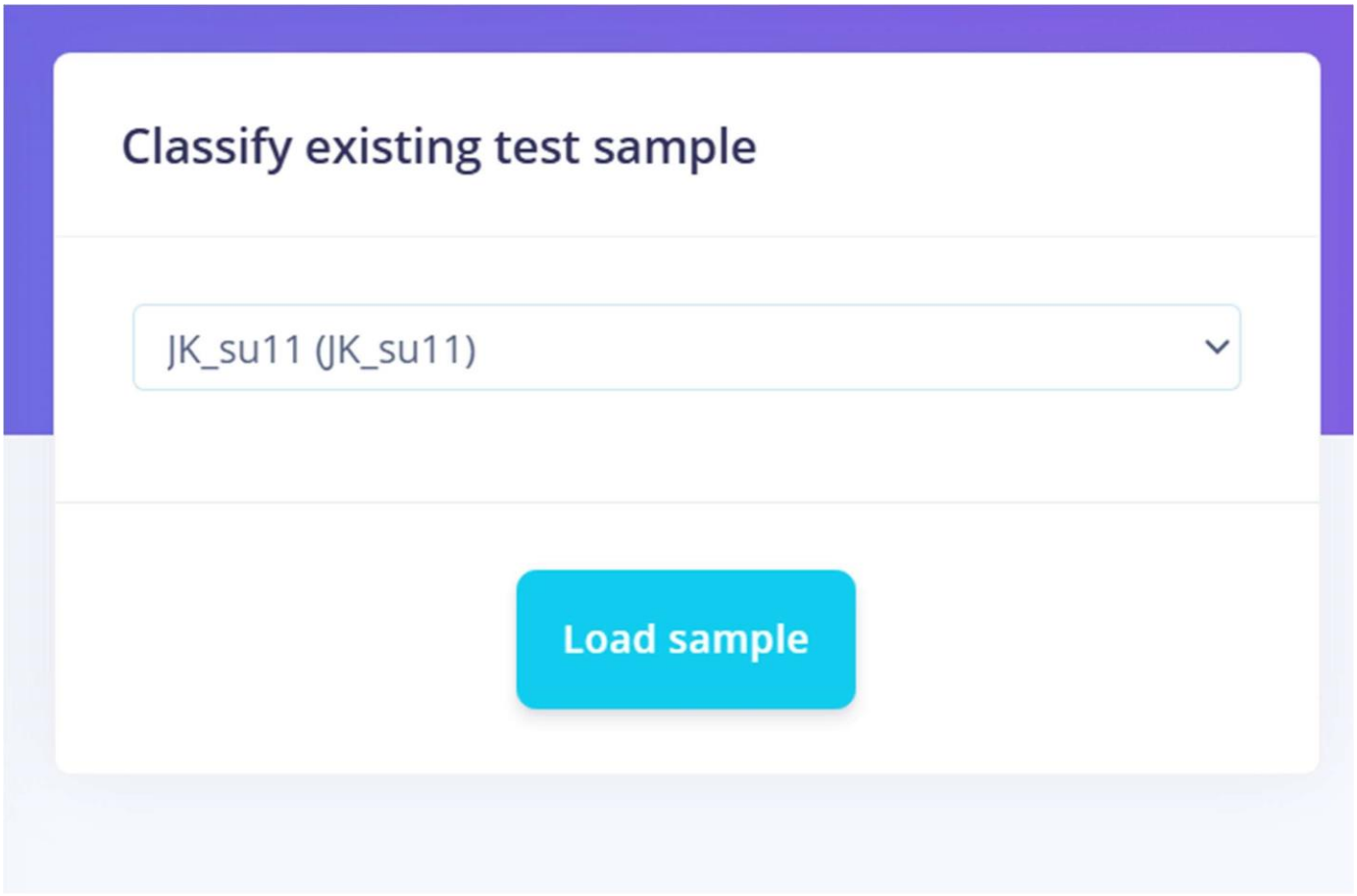
The Model Accuracy and Metrics-

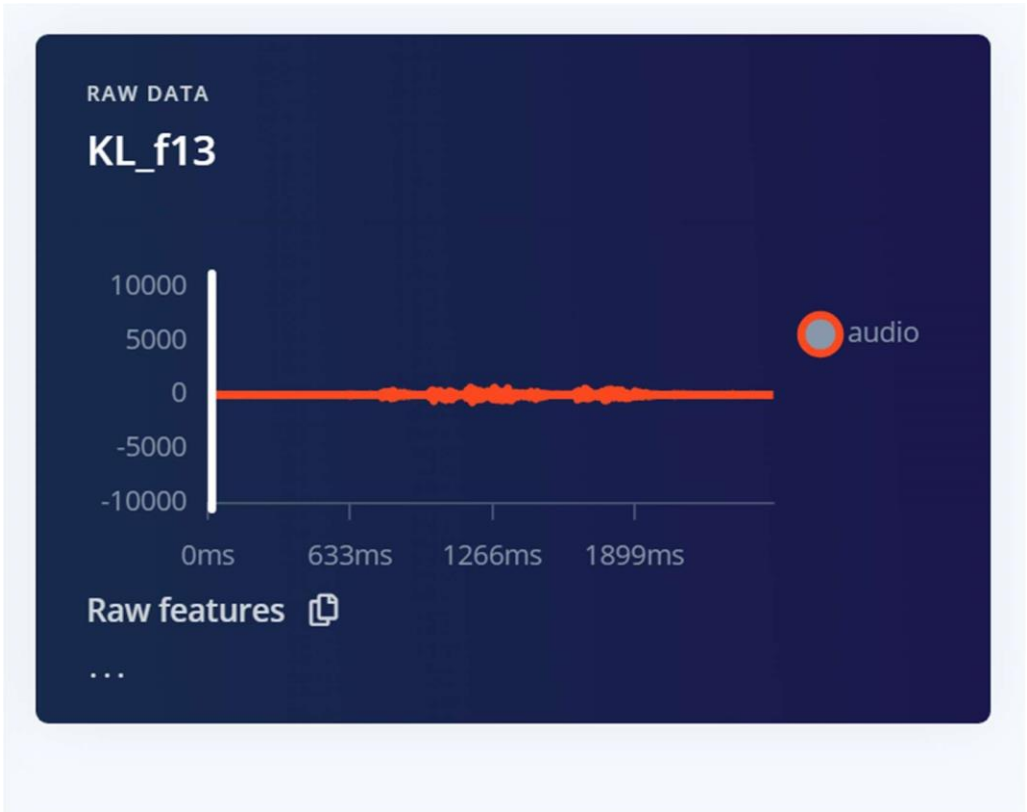


	precision	recall	f1-score	support
angry	0.86	0.71	0.78	45
disgust	0.54	0.98	0.70	48
fear	0.86	0.64	0.74	39
happy	0.74	0.74	0.74	54
neutral	0.82	0.69	0.75	39
sad	0.89	0.67	0.77	46
surprise	0.88	0.80	0.83	44
accuracy			0.75	315
macro avg	0.80	0.75	0.76	315
weighted avg	0.79	0.75	0.76	315

Output on Edge Impulse-

(deployed the .tflite file of the original model)





Classification result

Summary

Model version: Unoptimized (float32)

Name: KL_f13

Label: KL_f13

CATEGORY	COUNT
suprise	2
fear	18
sad	5
angry	1
uncertain	0