Smart system Final Project

Voice Emotion Recognition

Team members:

- Amr khaeld abdelhady hassan 2305224
- Abdalla ahmed fawzy 2305322
- Hamza ismail mahran 2305349

🎯 Objective of the project :

The goal of this project is to classify emotions (e.g., happy, sad, angry, neutral, etc.) from human speech using audio features. By analyzing the tone, pitch, intensity, and other vocal attributes, you can determine the emotional state of the speaker .

🛠️ Technologies Used

- **Python Libraries:** librosa, numpy, pandas, matplotlib, tensorflow/keras, streamlit
- **Model Type:** Deep Learning (LSTM)
- **UI Framework:** Streamlit

📚 Datasets Used

```python
# defining data
RAVDESS = "C:/Users/abdal/Desktop/Smart Final Project/Ravdess"
CREMA = "C:/Users/abdal/Desktop/Smart Final Project/Crema"
TESS = "C:/Users/abdal/Desktop/Smart Final Project/Tess"
SAVEE = "C:/Users/abdal/Desktop/Smart Final Project/Savee"

ravdess_dir_lis = os.listdir(RAVDESS)
path_list = []
gender_list = []
emotion_list = []

emotion_dic = {
    '03' : 'happy',
    '01' : 'neutral',
    '04' : 'sad',
    '05' : 'angry',
    '06' : 'fear',
    '07' : 'disgust',
}
```

[ ] Python

The system uses four popular public datasets for training:

1. **RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song):**
- Contains 24 actors speaking with 8 emotions.
- Format: .wav
- Gender and emotion encoded in filenames.
2. **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset):**
- Over 7,000 audio files with 6 emotions.
- Male and female speakers.
- High-intensity expressions are used.
3. **TESS (Toronto Emotional Speech Set):**
- 2 female actors reading 200 target words.
- Each word spoken in 7 different emotions.
4. **SAVEE (Surrey Audio-Visual Expressed Emotion):**
- 4 male speakers.
- 7 emotions recorded.

⚙️ Preprocessing Steps

- **Trimming and Padding:** Silence is trimmed, and the audio is padded to a fixed length.

```
def preprocess_audio(path):
    _, sr = librosa.load(path)
    raw_audio = AudioSegment.from_file(path)

    samples = np.array(raw_audio.get_array_of_samples(), dtype='float32')
    trimmed, _ = librosa.effects.trim(samples, top_db=25)
    padded = np.pad(trimmed, (0, 180000-len(trimmed)), 'constant')
    return padded, sr
```

- **Feature Extraction:**
  - **ZCR (Zero Crossing Rate)**
  - **RMS (Root Mean Square Energy)**
  - **MFCC (Mel Frequency Cepstral Coefficients)**

```
for row in df.itertuples(index=False):
    try:
        y, sr = preprocess_audio(row.path)

        zcr = librosa.feature.zero_crossing_rate(y, frame_length=FRAME_LENGTH, hop_length=HOP_LENGTH)
        rms = librosa.feature.rms(y=y, frame_length=FRAME_LENGTH, hop_length=HOP_LENGTH)
        mfccs = librosa.feature.mfcc(y=y, sr=sr, n_mfcc=13, hop_length=HOP_LENGTH)

        zcr_list.append(zcr)
        rms_list.append(rms)
        mfccs_list.append(mfccs)

        emotion_list.append(encode(row.emotion))
    except:
        print(f"Failed for path: {row.path}")
```

- **Encoding Emotions:** Each emotion is mapped to an integer class.

```
emotion_dic = {
    'neutral' : 0,
    'happy'   : 1,
    'sad'     : 2,
    'angry'   : 3,
    'fear'    : 4,
    'disgust' : 5
}
```

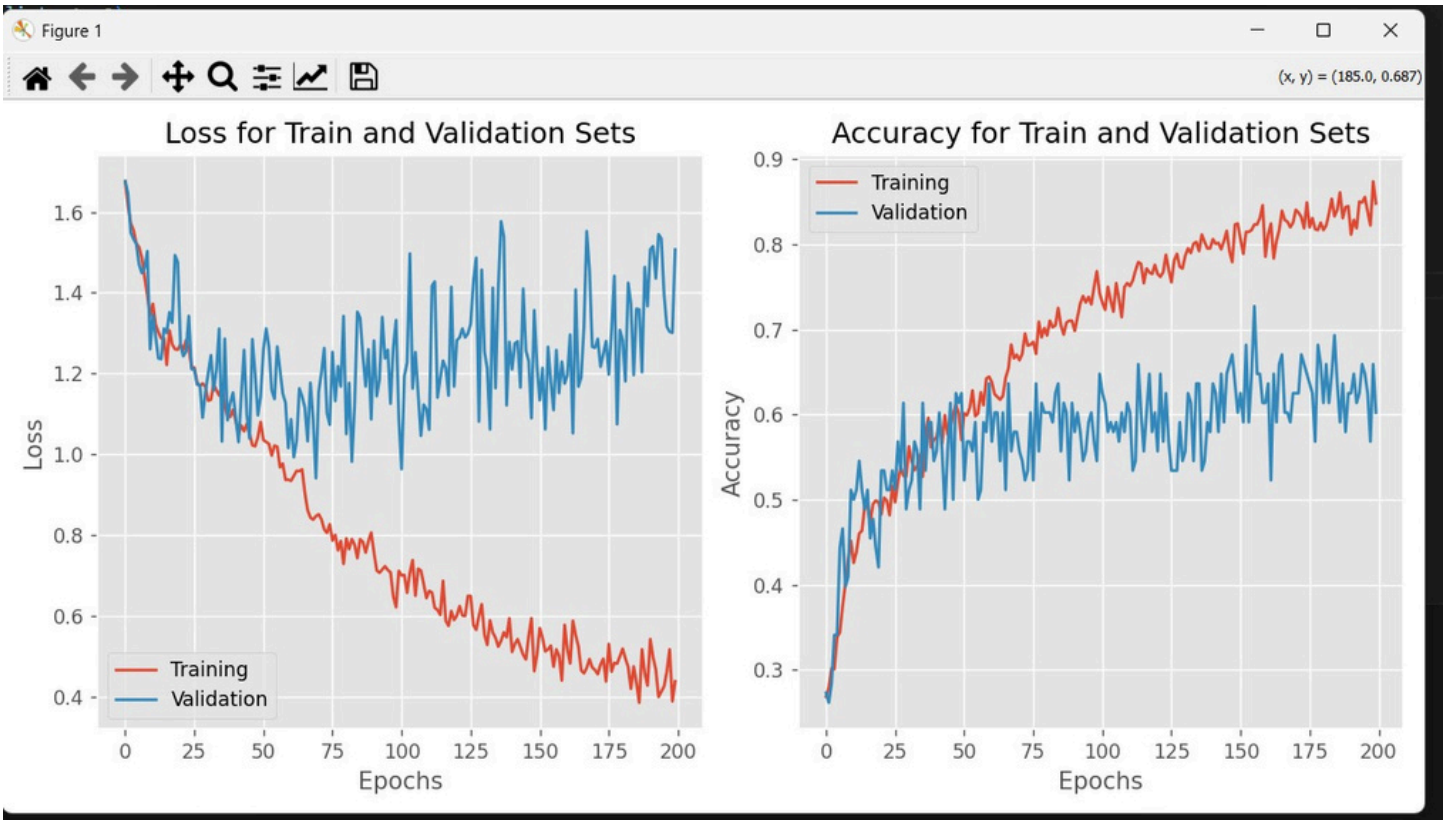
The raw audio is first trimmed to remove silence and then padded to a uniform length. For each audio clip, features such as Zero Crossing Rate (ZCR), Root Mean Square Energy (RMS), and 13 Mel Frequency Cepstral Coefficients (MFCCs) are extracted using Librosa. These features capture both time-domain and frequency-domain characteristics of the speech. Finally, string-based emotion labels are converted into numeric form for model training.

🧠 Model Architecture

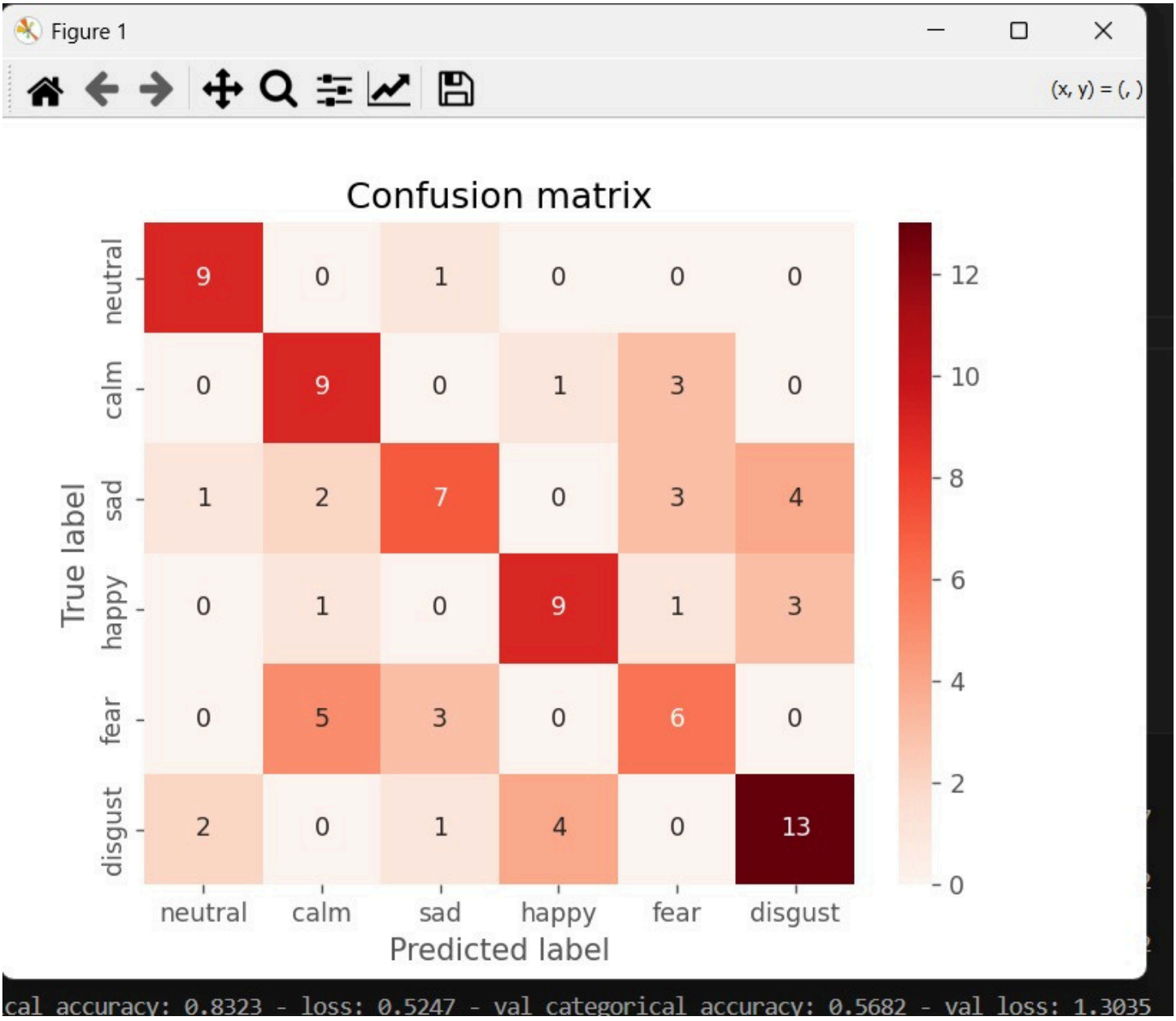
- **Model Type:** LSTM-based sequential neural network.
- **Layers:**
  - LSTM (64 units) × 2
  - Dense (Softmax activation)
- **Loss Function:** Categorical Crossentropy
- **Optimizer:** RMSProp
- **Evaluation Metrics:** Accuracy

📈 Performance

- **Training and validation loss/accuracy are visualized over epochs.**



- **Confusion matrix** is used to evaluate performance on validation data



🎯 Emotions Detected

The model is trained to recognize the following 6 emotions:

- Neutral
- Happy
- Sad
- Angry
- Fear
- Disgust

✅ Conclusion

This system demonstrates the potential of combining audio signal processing with deep learning to build real-time emotion recognition tools. Such applications can be useful in healthcare, customer service, entertainment, and human-computer interaction.