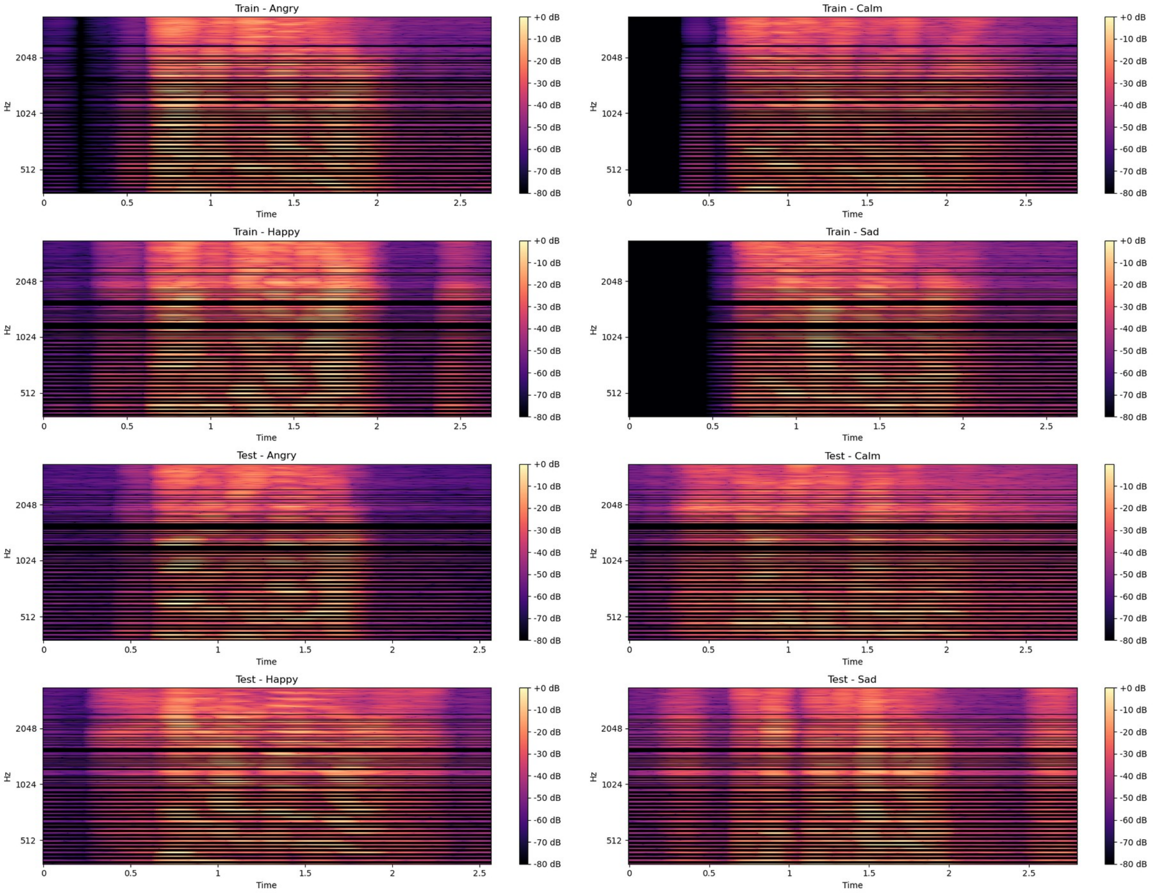
SIT789 – Robotics, Computer Vision and Speech Processing

Distinction Task 8.2: Speech emotion recognition using spectral features and deep learning approaches

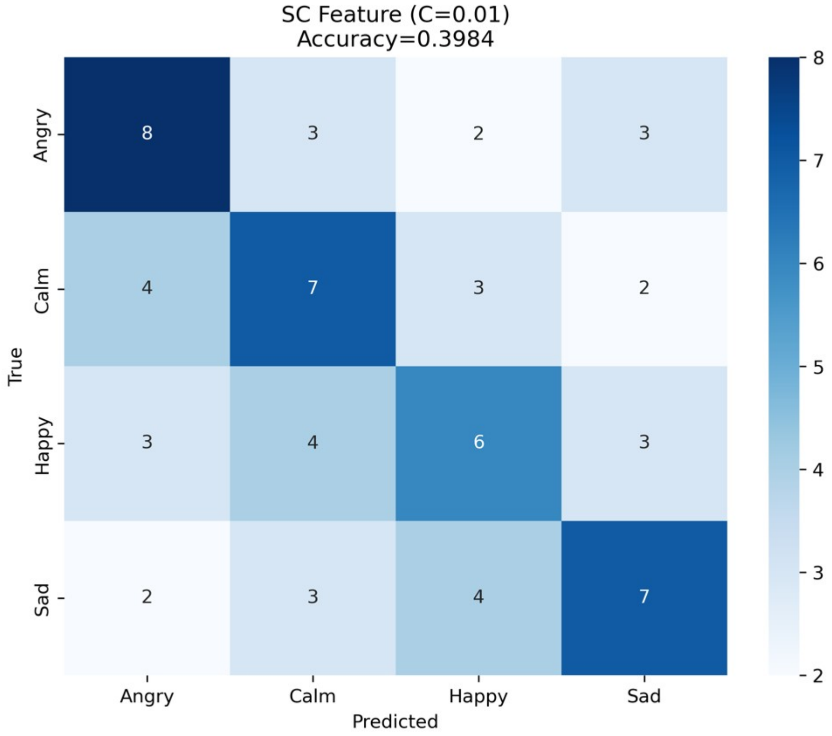
a. **Visualization of mel-scale spectrograms for training and test audio clips, ensuring that there is at least one clip per emotion class for both training and testing (Section 1)**



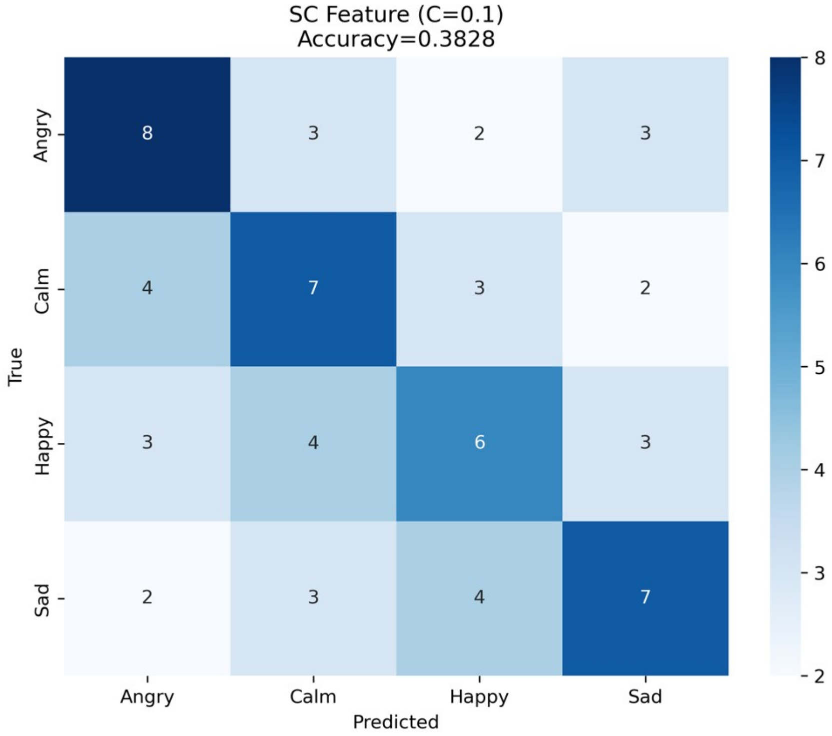
b. **Overall recognition accuracy and confusion matrix for each spectral feature type (SC, SBW, SBE) using your SVM model (Section 2), as well as identifying the spectral feature type with the highest overall recognition accuracy.**

Evaluating SVM with SC features

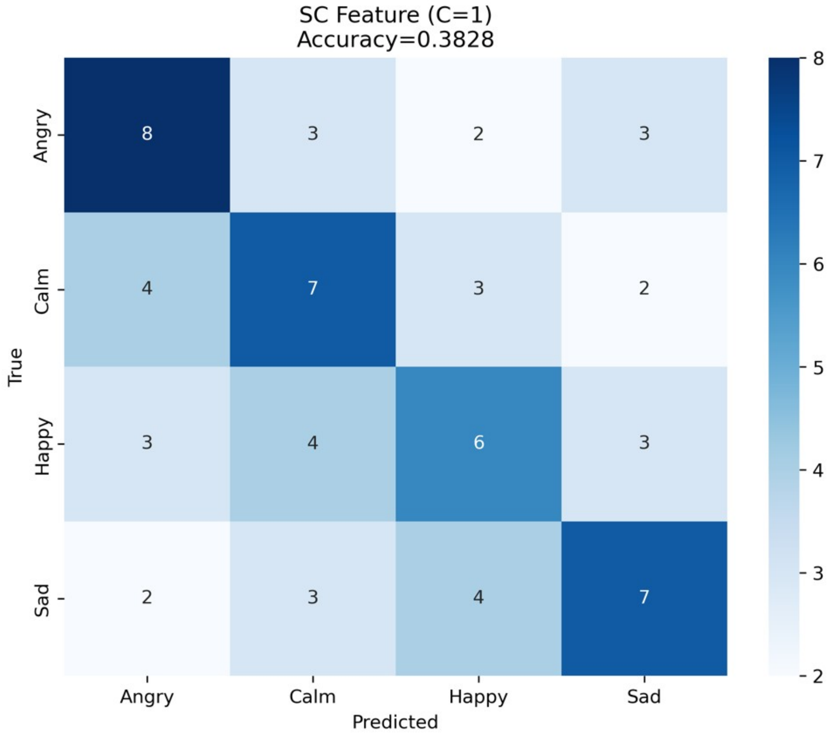
C = 0.01, Accuracy: 0.3984



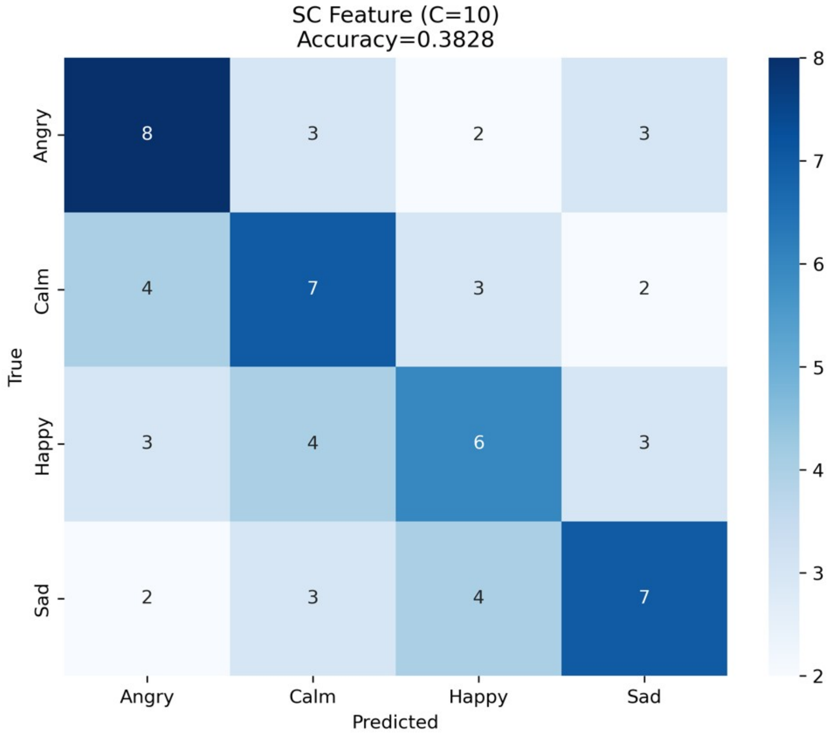
C = 0.1, Accuracy: 0.3828



C = 1, Accuracy: 0.3828

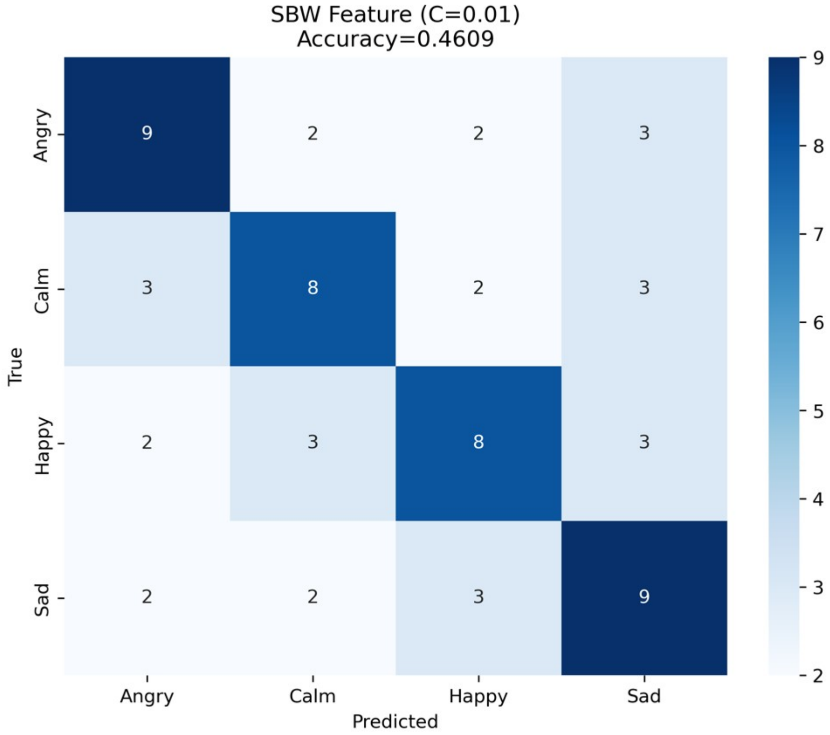


C = 10, Accuracy: 0.3828

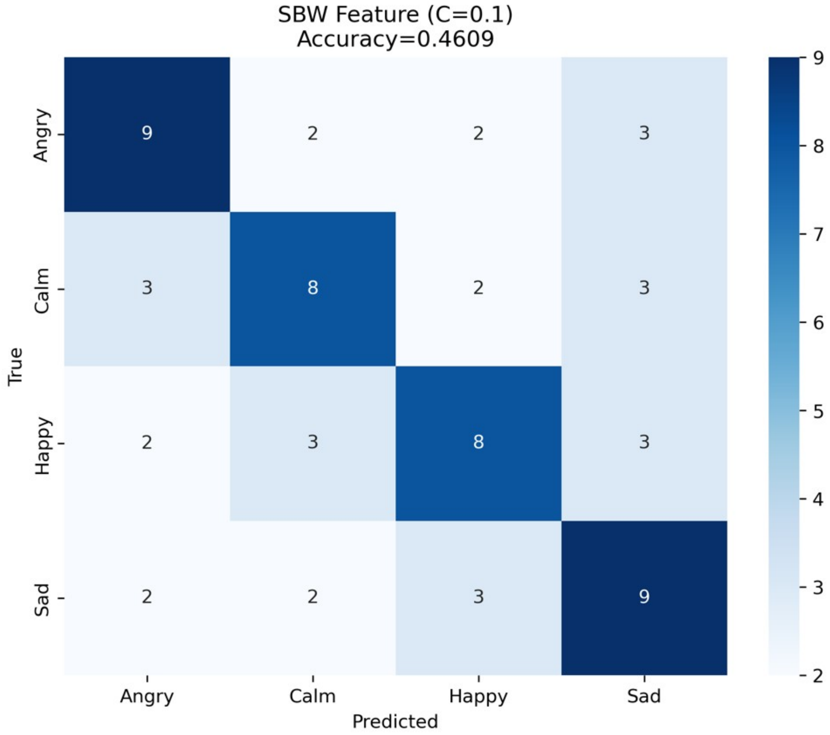


Evaluating SVM with SBW features

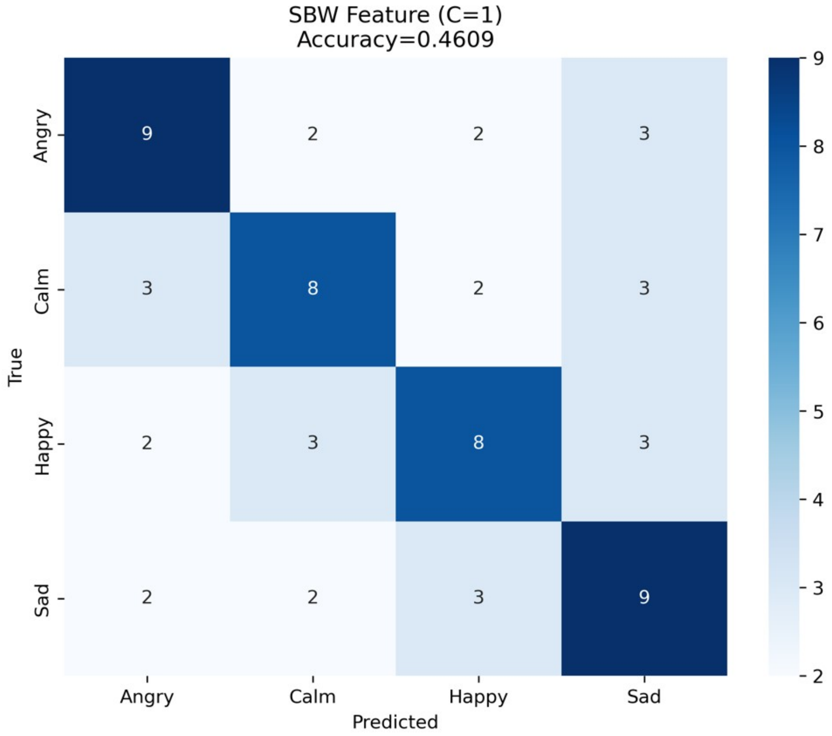
C = 0.01, Accuracy: 0.4609



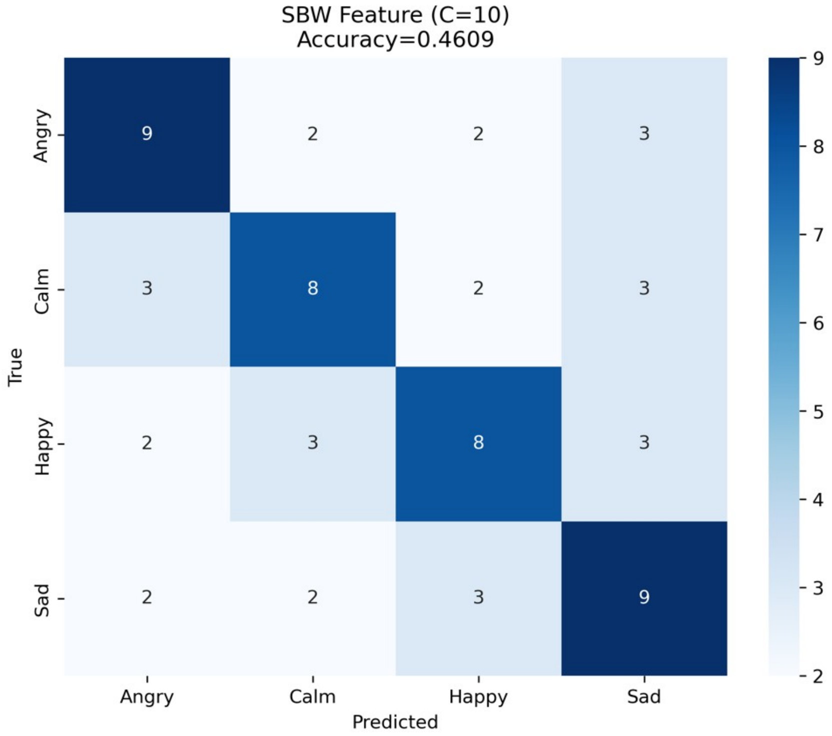
C = 0.1, Accuracy: 0.4609



C = 1, Accuracy: 0.4609

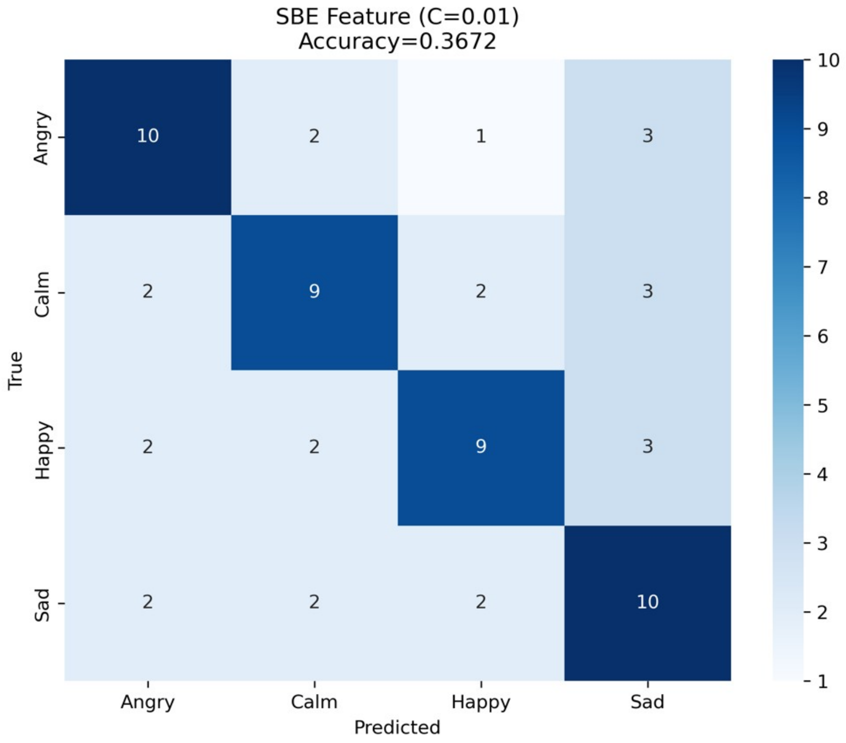


C = 10, Accuracy: 0.4609

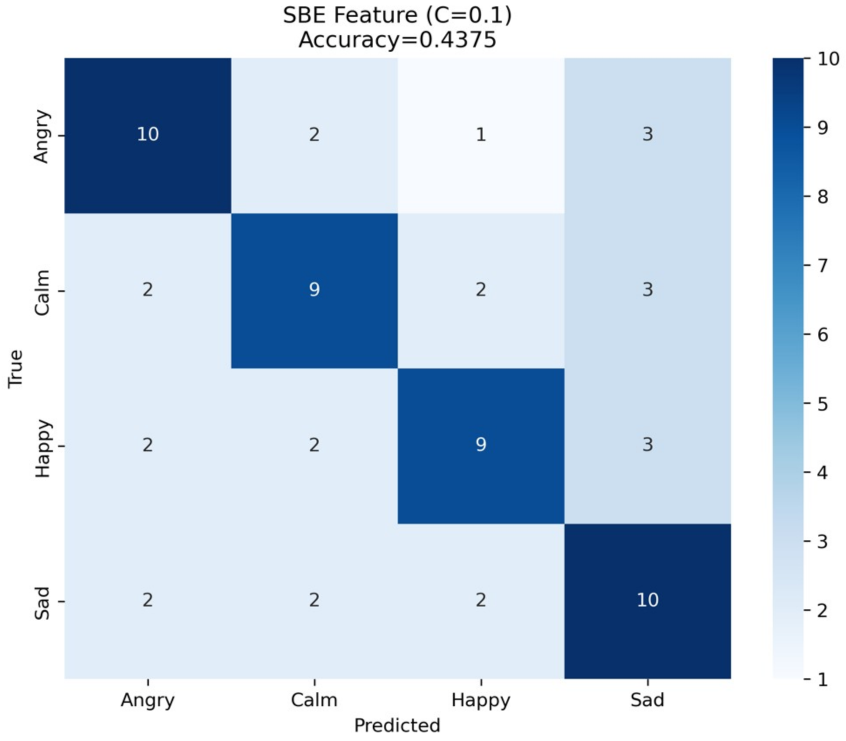


Evaluating SVM with SBE features

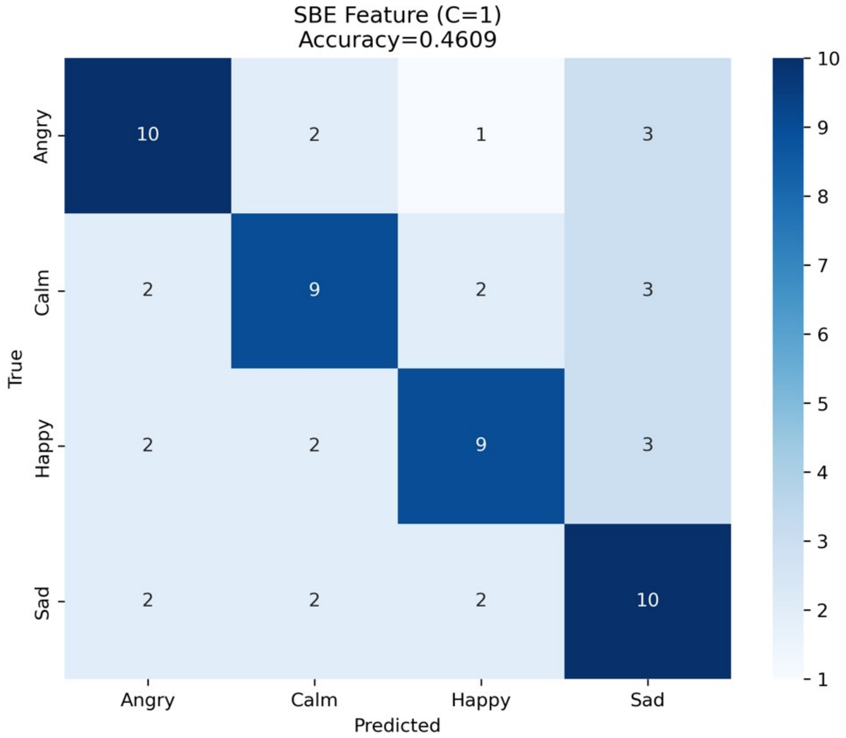
C = 0.01, Accuracy: 0.3672



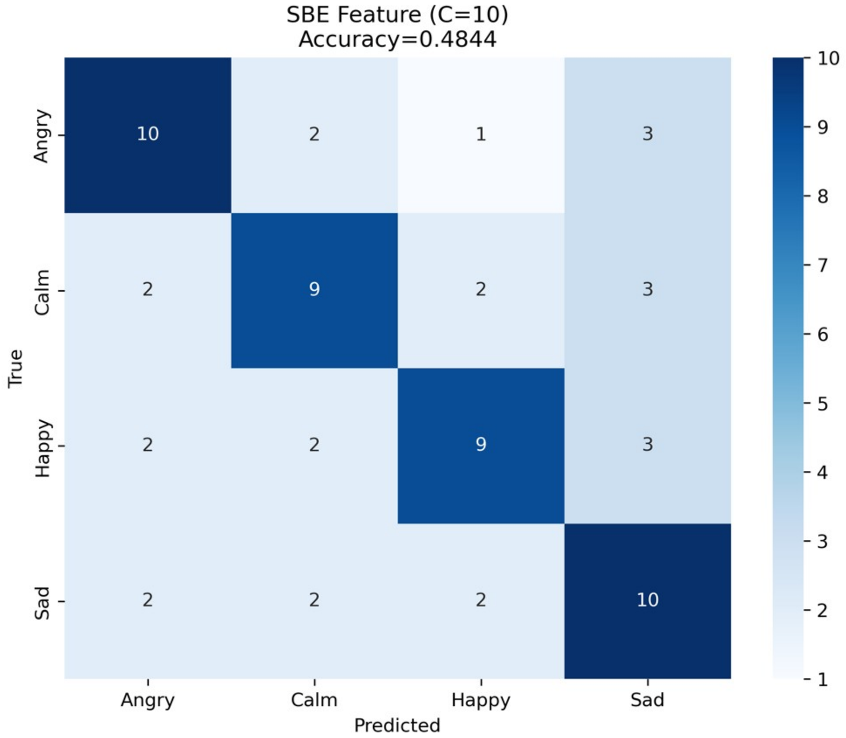
C = 0.1, Accuracy: 0.4375



C = 1, Accuracy: 0.4609



C = 10, Accuracy: 0.4844



BEST PERFORMANCE:

- Feature Type: SBE (Spectral Band Energy)
- C value: 10
- Accuracy: 0.4844 (48.44%) Comment: SBE features with C=10 performed best, suggesting that spectral band energy characteristics are most effective at distinguishing between different emotions in speech.

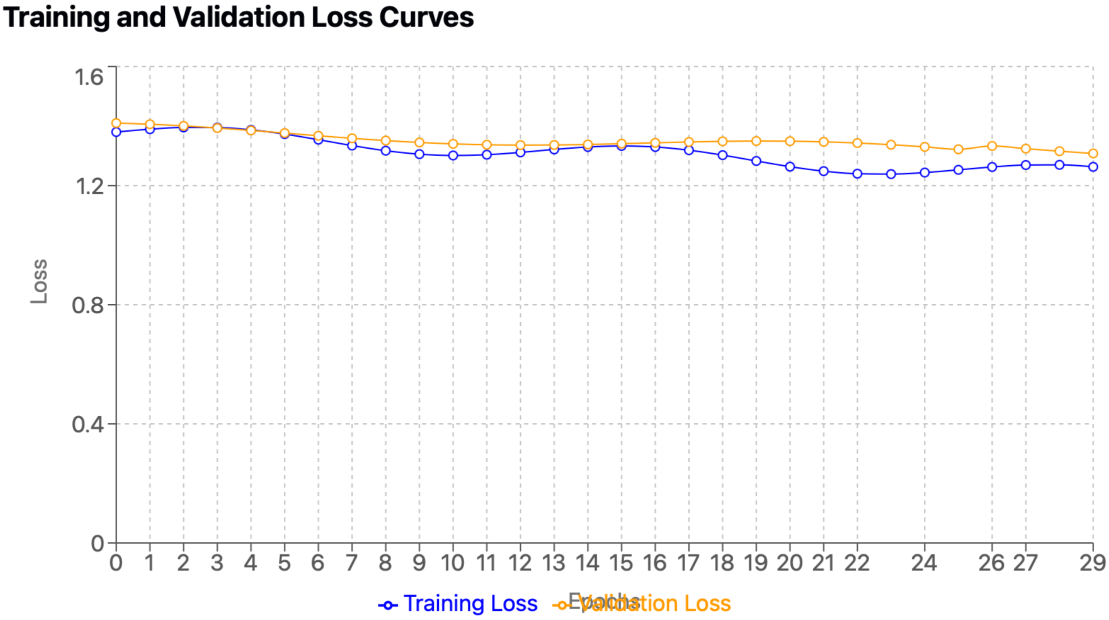
WORST PERFORMANCE:

- Feature Type: SBE
- C value: 0.01
- Accuracy: 0.3672 (36.72%) Comment: The same feature type (SBE) with a much lower C value performed worst, indicating that SBE features require a stronger regularization parameter (higher C) to be effective.

Alternative Worst:

- Feature Type: SC (Spectral Centroid)
- C values: 0.1, 1, and 10 all had
- Accuracy: 0.3828 (38.28%) Comment: Spectral Centroid features showed consistently poor performance across multiple C values, suggesting this feature might not be as discriminative for emotion recognition.

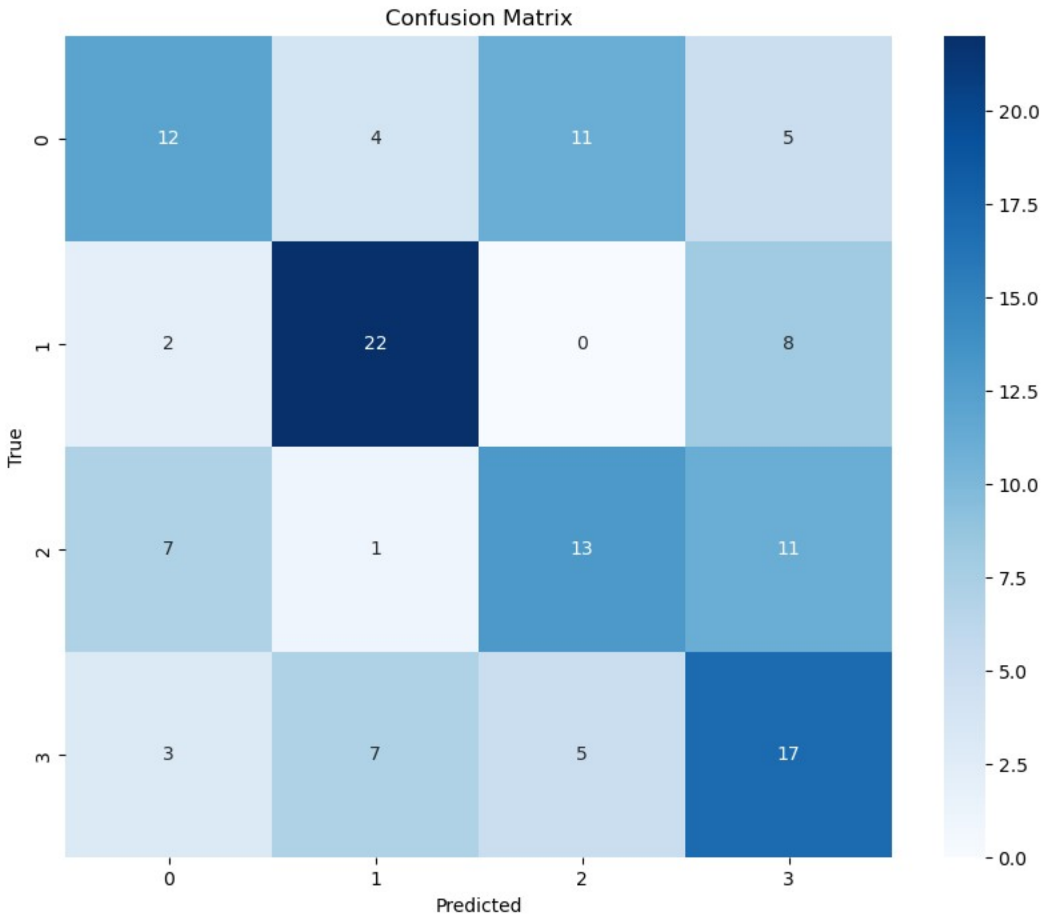
c. **A Learning curve (or loss curve) generated from the training of your 1DCNN model (Section 3.1, see Task 5.1).**



d. **Overall recognition accuracy and confusion matrix for your 1D CNN model (Section 3.1).**

Test set: Average loss: 1.1466
Test set accuracy: 57.81%

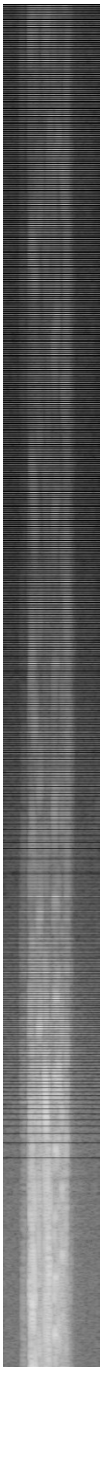
Final model accuracy: 57.81%



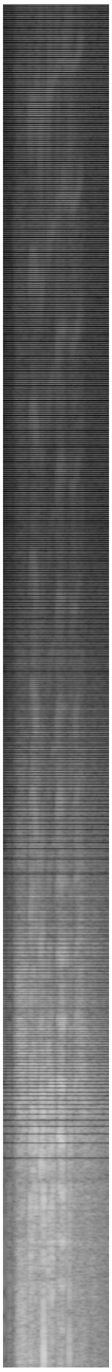
e. **Mel-scale spectrogram images generated by the spectrogram2image method on training and test audio clips, ensuring at least one clip per training and testing for each emotion class (Section 3.2).**

Test set:

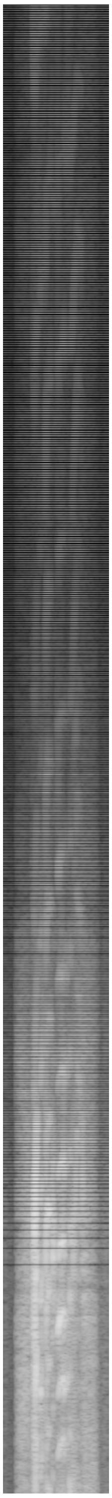
Angry:



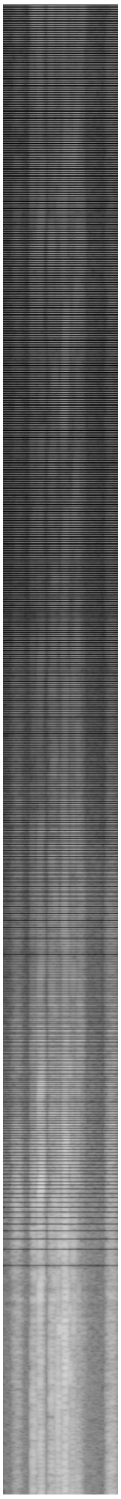
Calm:



Happy:

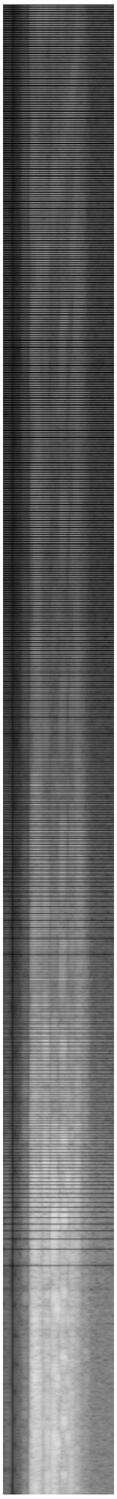


Sad:

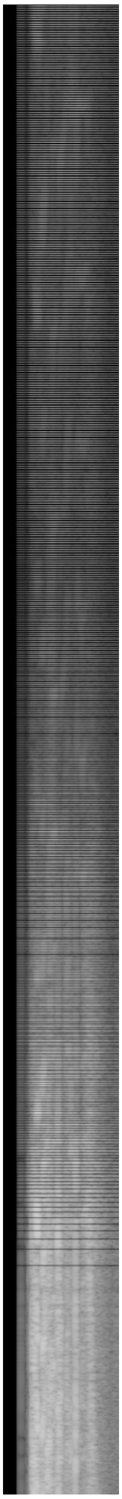


Train set:

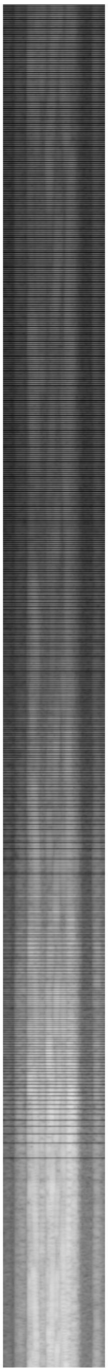
Angry:



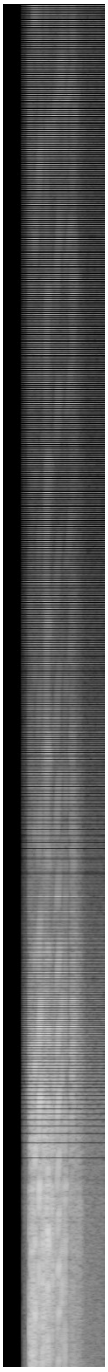
Calm:



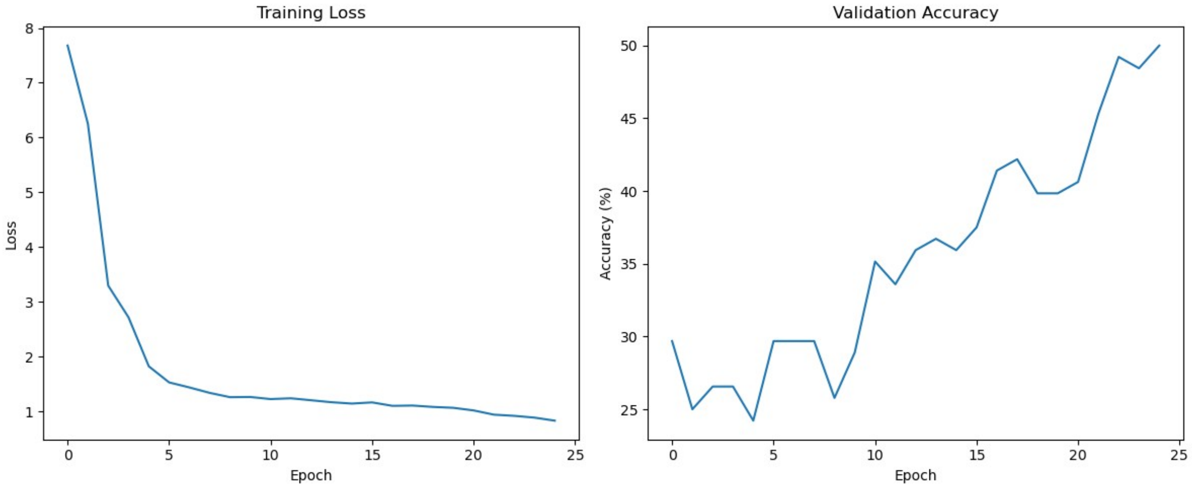
Happy:

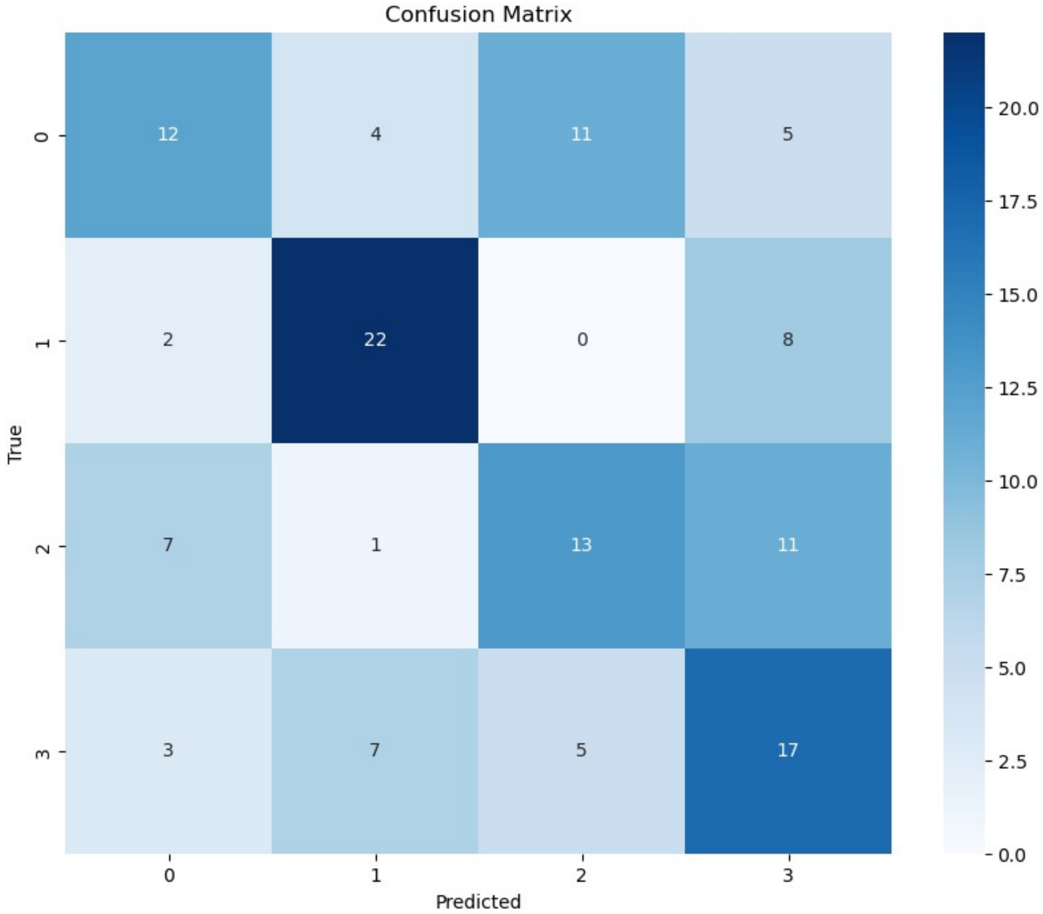


Sad:



f. **Learning curve (or loss curve) generated from the training of your 2D CNN model (Section 3.2, see Task 5.1).**





g. Overall recognition accuracy and confusion matrix for your 2D CNN model (Section 3.2).

Final Test Accuracy: 50.00%

h. Your opinion on data augmentation (Section 3.2).

Augmenting data for spectrograms in speech emotion recognition demands meticulous consideration, as conventional image augmentation techniques may not be suitable. Techniques like cropping, flipping, and rotating, though prevalent in 2D CNNs, may actually degrade model performance for spectrograms. This is because spectrograms depict specific time-frequency relationships, with the x-axis representing time progression, the y-axis indicating frequency, and intensity showing energy at each point. Horizontal flipping would reverse temporal sequences, vertical flipping would invert frequency relationships, and random cropping might eliminate critical emotional cues.

Similarly, rotation would disrupt both time and frequency dimensions, potentially destroying meaningful patterns. Instead, more suitable augmentation techniques for speech emotion data might involve adding minor amounts of noise, slight time-stretching/compression, small pitch shifts, or time/frequency masking. These methods retain the acoustic and emotional characteristics of speech.

Thus, while data augmentation can be advantageous for increasing training data volume, it is essential to apply domain-specific augmentation methods that preserve the inherent time-frequency structure of spectrograms rather than using standard image augmentation techniques.