TinyML Device for Emotion Detection

Yuancheng Cao
yuc094@ucsd.edu
University of California, San Diego
San Diego, California, USA

Abstract

This project aims to develop an emotion-sensing device that analyzes speech using machine learning models and a microphone to detect and response to a person's emotional state in real-time. The device provides visual feedback through colored LED lights, creating an immersive environment that reflects the user's emotions. Experimental evaluation and model testing conducted in Edge Impulse revealed the presence of biases in the training data and the need for improvement in detecting certain emotions, particularly happiness and neutral speech. The overall accuracy of the model was found to be 51.28%, with the model struggling to classify 56.4% of happy audio files and 61.9% of neutral ones, while also confusing happiness with anger in nearly 15% of the cases. Future work should focus on curating and expanding the training dataset, experimenting with preprocessing techniques and model architectures, and integrating a music recommendation feature based on the detected emotions to enhance the user experience and promote emotional well-being.

1 Introduction

In recent years, mental health has become a growing concern, particularly among college students. According to statistics, in 2023, over 76% of college students experienced moderate to serious psychological distress, with more than 9 in 10 students stating that academic challenges affect their mental well-being [1].Furthermore, in 2022, 25% of students reported feeling isolated from others, and only 40% believed that their school was doing enough to support student mental health [2, 3]. These alarming figures highlight the need for innovative solutions to monitor and address the emotional well-being of individuals.

Emotion recognition has been a topic of interest in various fields, including psychology, computer science, and human-computer interaction. Traditional methods of emotion detection often rely on self-reporting or human observation, which can be subjective and time-consuming. However, with the advancement of machine learning techniques and the increasing availability of small, powerful devices, it is now possible to develop automated emotion recognition systems that can analyze speech and provide real-time feedback.

The goal of this project is to create a device that can accurately detect and respond to a person's emotional state by analyzing their speech using Arduino Nano 33 BLE and a convolutional neural network (CNN) model. The device will provide a visual representation of the detected emotions through colored LED lights, creating an immersive and responsive environment. For instance, if the device detects happiness in the person's voice, the lights may turn green, while a sad tone could trigger blue lights. By offering a simple and intuitive way to monitor emotional well-being, this device aims to promote mental health awareness and provide support to individuals who may be experiencing psychological distress.

The proposed emotion-sensing device has the potential to be particularly beneficial in settings such as college campuses, where students often face significant mental health challenges. By creating a room that responds to emotions and moods, the device can help individuals track their mental state and alert others to their need for support. This can foster a more empathetic and supportive environment, encouraging open discussions about mental health and reducing the stigma surrounding psychological distress.

2 Technical Aspects

2.1 Design Choices

Edge Impulse was selected to build the emotion-sensing device. Initially, audio recordings of different emotions were uploaded and labeled in Edge Impulse. A machine learning model was then trained to recognize the emotions based on the labeled data. The model was tested using audio samples from the testing set to ensure its accuracy. Finally, the code was obtained to deploy the model onto the Arduino Nano 33 BLE. Section 2.2, 2.3, 2.4, 2.5, 2.6, 2.7 will provide a detailed explanation of these processes.

For the hardware, an Arduino Nano 33 BLLE was chosen. It was programmed to use different colored light to represent each emotion: red for anger, green for happiness, yellow for neutrality, and blue for sadness. The device was set up to record new aduio every 5 seconds, with a 2-second interval to prepare between recordings. The final outcome is capable of listening to speech, determining the speaker's emotion using the machine learning model, and displaying the corresponding emotion through a colored light.

2.2 Data Preparation

The data used in this project is taken from the Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D) available on Kaggle. The dataset consists of 7,442 original clips contributed by 91 actors, including 48 males and 43 females, with ages ranging from 20 to 74 years old. The actors represent various racial backgrounds, including African American, Asian, White, Hispanic, and Unknown. Each actor in the dataset delivers speech from a selection of 12 sentences. Additionally, each emotional state is expressed at four different levels: low, medium, high, and unknown, providing a nuanced understanding of emotional expression.

For this project, I focused on a subset of the dataset, utilizing 3,537 clips. This subset includes clips categorized under the emotions Anger, Happy, Neutral, and Sad. By narrowing the scope to these four emotions, the project aims to create a more focused and manageable dataset for emotion detection tasks.

2.3 Impulse Design: Classification Model

After preparing the data, the next step was to design the machine learning model in Edge Impulse. A classifcation model was chosen

to categorize audio recordings into different emotion categories, specifically angry, happy, neutral, and sad. The Mel-Frequency Cepstral Coefficients (MFCC) feature was selected for the audio data, as it effectively captures the essential components of the audio signal that are relevant to emotional expression.

The audio clips in the dataset have a minimum duration of 2 seconds. To ensure that the model can process these clips effectively, the window size for each audio segment was set to 2 seconds. This means that the model will analyze the audio data in 2-second intervals, aligning with the minimum length of the audio clips in the dataset. The audio frequency was set to 16,000 Hz, which is sufficient for capturing the necessary information from the audio signal.

By setting the window size to match the minimum duration of the audio clips, the model can efficiently process the entire dataset without any loss of information. This design choice ensures that the model has access to the complete audio clip for each sample, enabling it to make accurate predictions on the emotional state represented in the clip. The selection of the MFCC feature is based on its proven effectiveness in speech recognition and emotion classification tasks. MFCCs capture the spectral envelope of the audio signal, which contains information about the vocal tract characteristics and emotional cues.

The classification model designed in Edge Impulse takes advantage of the minimum 2-second duration of the audio clips in the dataset. By setting the window size to 2 seconds and utilizing the MFCC feature, the model can process the audio data efficiently and accurately classify the emotions represented in each clip. This design approach ensures that the model can fully utilize the available information in the dataset and provide reliable predictions on the emotional state of the speaker.

2.4 Audio (MFCC)

The MFCC feature graph for the audio recordings reveals complex patterns and characteristics of the sound data. As shown in Figure 1, the graph provides a visual representation of how different components of the sound evolve over time. Consistent red or blue regions in the graph indicate that certain sound features remain stable for a period, suggesting that these features are potentially important for distinguishing between different emotional states.

On the other hand, areas with rapid changes in color and intensity signify that the sound features are undergoing swift transitions. These dynamic regions in the graph may correspond to the more expressive or emotive parts of the audio recordings, where the speaker's emotional state is more pronounced. By analyzing these patterns and variations in the MFCC feature graph, valuable insights can be gained into the relationship between specific sound characteristics and the expressed emotions.

2.5 Training 1D Convolutional Model

For the emotion recognition task using audio data, a 1D Convolutional Neural Network (CNN) model was chosen. 1D CNNs are well-suited for processing sequential data, such as audio signals, as they can effectively capture local patterns and temporal dependencies. The neural network architecture for the 1D CNN model was configured as follows:

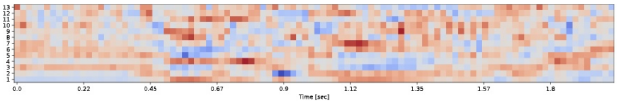

Figure 1: Digital Signal Processing (DSP) – Cepstral Coefficients

(1) Input layer: The input layer consists of 2,528 features, which correspond to the MFCC features extracted from the audio recordings.
(2) Reshape layer: The reshape layer is used to reorganize the input features into a suitable shape for the subsequent convolutional layers. In this case, the features are reshaped into 32 columns.
(3) 1D conv / pool layers: The model includes three sets of 1D convolutional and pooling layers. Each set consists of a convolutional layer followed by a pooling layer. The number of neurons in each convolutional layer increases progressively (8, 16, and 32), allowing the model to learn increasingly complex features. The kernel size for each convolutional layer is set to 3, which determines the size of the convolutional filters. The number of layers in each set is 1.
(4) Flatten layer: After the convolutional and pooling layers, the flatten layer is used to convert the 2D feature maps into a 1D feature vector. This step prepares the data for the final classification layers.
(5) Dropout layer: A dropout layer is added with a rate of 0.5, meaning that 50% of the neurons are randomly dropped out during training.

2.5.1 Iteration 1 and 2. The training process for the emotion recognition mode was conducted in two rounds, referred to as Iteration 1 and Iteration 2. In both iterations, the model underwent 100 training cycles with a learning rate of 0.005. The dataset was split into a training set and a testing set, with 80% of the data used for training and the remaining 20% for testing. Additionally, 20% of the training set was allocated for validation purposes.

Iteration 1 yielded an accuracy of 58.2% (see Figure 2), which was deemed unsatisfactory. To address this, the dataset was revisited, and certain audio recordings that were potentially confusing to the model were relabeled. These modifications aimed to provide clearer distinctions between the different emotional states. With the updated dataset, Iteration 2 was initiated, resulting in an improved model accuracy of 59.2% (see Figure 3). This increase in accuracy suggests that the relabeling effort helped the model better differentiate between the emotions present in the audio recordings.

2.5.2 Iteration 3. Following the completion of Iteration 2, the model's accuracy was still not satisfactory. To further improve the performance, the dataset was revisited once again. A thorough examination of the audio recordings revealed instances that were potentially mislabeled or lacked clarity. These recordings were removed from the dataset, and some other recordings were relabeled to ensure better categorization.

With the dataset refined, Iteration 3 of the training process was initiated. The same settings were maintained, including 100 training

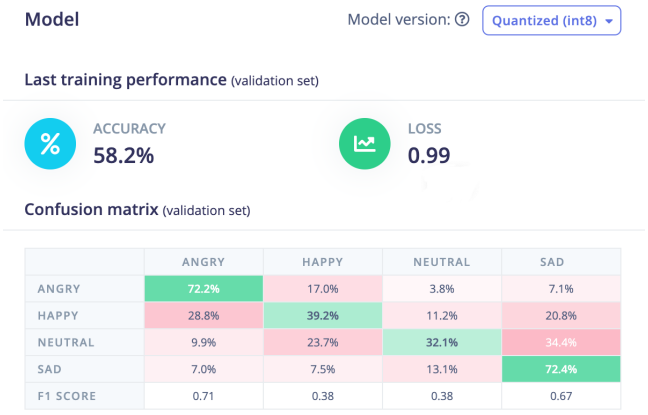

Figure 2: Accuracy and Confusion Matrix of the First Iteration of the 1D Convolutional Model

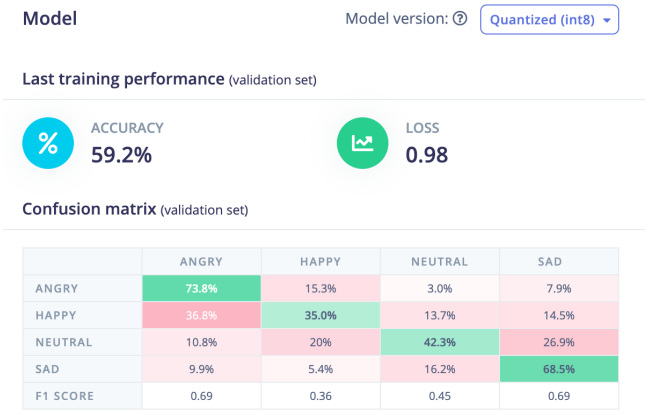

Figure 3: Accuracy and Confusion Matrix of the Second Iteration of the 1D Convolutional Model

cycles, a learning rate of 0.005, and splitting the data into 80% for training and 20% for testing, with 20% of the training set used for validation. The result of Iteration 3 yielded an accuracy of 69.5% (see Figure 4), demonstrating a significant improvement compared to the previous iterations. Fixing the dataset by removing confusing recordings and relabeling others helped the model classify emotions in the audio data better.

2.6 Training 2D Convolutional Model

Following the improvement observed in Iteration 3, a different approach was explored to potentially further enhance the accuracy. A 2D Convolutional Model was employed for this purpose. The choice of a 2D CNN was motivated by its ability to capture spatial relationships and patterns within the audio data.

The main difference between a 1D CNN and a 2D CNN lies in the dimensionality of the convolution operation. In a 1D CNN, the convolution operates along a single dimension, typically the

Figure 4: Accuracy and Confusion Matrix of the Third Iteration of the 1D Convolutional Model

temporal dimension of the audio signal. This makes 1D CNNs well-suited for processing sequential data and capturing temporal dependencies. On the other hand, a 2D CNN performs convolution along both the temporal and frequency dimensions of the audio spectrogram. By considering the relationships between adjacent time frames and frequency bins, a 2D CNN can potentially extract more complex features from the audio data.

The architecture of the 2D CNN model used in Iteration 4 was structured as follows: The input layer consists of 2,528 features, which are then reshaped into 32 columns. The first 2D convolutional and pooling layer has 8 filters with a kernel size of 3, followed by a dropout layer with a rate of 0.5 to prevent overfitting. The second 2D convolutional and pooling layer has 16 filters with a kernel size of 3, again followed by a dropout layer. Finally, the output of the convolutional layers is flattened before being passed to the final classification layers.

2.6.1 Iteration 4. The training parameters were set to 100 training cycles, a learning rate of 0.005, and the data was split into 80% for training and 20% for testing, with 20% of the training set used for validation. The new model was set up, and Iteration 4 of the training process was initiated. Upon completion, the accuracy was evaluated, and it was found to be 65.8% (see Figure 5). However, this accuracy was lower compared to the 69.5% achieved by the previous model in Iteration 3. After analyzing the results, the decision was made to proceed with the model from Iteration 3, as it demonstrated the highest accuracy among all the iterations conducted.

Another reason why iteration 3's model is better than iteration 4's model, as shown in Figure 6, Most data points overlap, indicating significant misclassifications. Only the "sad" and "angry" emotions show better separation and higher correct classification rates. The model struggles with "happy" and "neutral" emotions, with many incorrect predictions. Using the iteration 3 model, which likely performed better, is suggested to improve accuracy.

Figure 5: Accuracy and Confusion Matrix of the 2D Convolutional Model

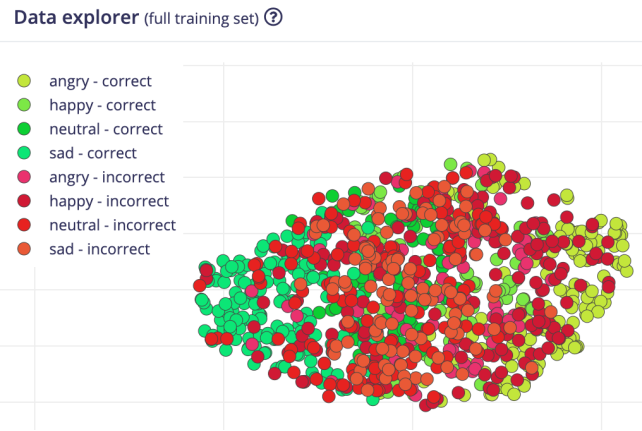

Figure 6: Data Explorer of the 2D Convolutional Model

2.7 EON Tuner

To further optimize the emotion-sensing model, EON Tuner was employed to explore potential improvements in performance. During the tuning process, EON Tuner tested 30 different architectures, including 29 variations of 1D convolutional models and one 2D convolutional model. After evaluating these architectures, the best model found by EON Tuner achieved an overall accuracy of 69%, which was comparable to the model obtained from Iteration 3.

An analysis of the best model's performance revealed that it excelled in recognizing angry and sad emotions, with an impressive accuracy of 79% for both categories. However, the model struggled somewhat with neutral emotions, correctly identifying them 63% of the time. The most challenging emotion for the model was happiness, which it correctly classified only half the time (see Figure 7). Nevertheless, the current model's performance was deemed satisfactory, and it was ready to be integrated into the emotion-sensing device.

Figure 7: Result of Best EON Tuner Model

3 Results and Analysis

3.1 Experimental

To assess the performance of the newly developed emotion-sensing device, an experimental evaluation was conducted to compare its accuracy against other existing methods. Two sentences were selected to test the device's ability to correctly identify the underlying emotions.

The first sentence, "Everything feels so heavy," was chosen, which intuitively conveyed a sense of sadness. To establish a benchmark, the sentence was analyzed using Amazon Web Services (AWS) Emotion Analysis and the Emotion Detection GPT tool. Interestingly, both of these tools classified the sentence as expressing anger rather than sadness. When the same sentence was tested using the TinyML-based emotion-sensing device, it successfully identified the emotion as sad, but only under specific conditions. The device required the sentence to be spoken in a very quiet manner, necessitating approximately 10 attempts to achieve accurate recognition. When the sentence was spoken at a normal volume, the device struggled to correctly classify the emotion.

To further evaluate the device'ss performance, a second sentence, "I am so happy today!" was used. This sentence clearly conveyed a sense of happiness. Both the AWS Emotion Analysis and the Emotion Detection GPT tool accurately identified the sentence as expressing happiness. However, TinyML-based emotion-sensing device encountered difficulties with this sentence, incorrectly classifying it as angry. In short, the experimental evaluation revealed the challenges associated with emotion detection and the potential for inconsistencies among different tools and methods.

3.2 Model Testing in Edge Impulse

After conducting initial tests with a few sentences, a comprehensive evaluation of the emotion-sensing device was performed using Edge Impulse. The overall accuracy of the model was found to be 51.28%, indicating that it correctly identified emotions approximately half the time. However, upon closer examination, it became apparent that the model struggled with certain emotions. As shown in Figure 8, it failed to classify 56.4% of the happy audio files and 61.9% of the neutral ones, demonstrating a high level of uncertainty in labeling these emotions.

Another significant issue that emerged during testing was the model's tendency to confuse happiness with anger. In nearly 15% of the cases, the model incorrectly labeled happy audio files as angry, representing a substantial misclassification error.

Further analysis of the results revealed that the model's emotion detection performance was unbalanced. It showed better accuracy in recognizing certain emotions compared to others, and often encountered confusion between specific pairs of emotions. This imbalance suggests that the model may have learned biases or patterns that allow it to detect certain emotions more accurately when struggling with others.

These findings highlight the need to further refine and optimize emotion perception models. Potential areas of improvement include collecting more diverse and balanced datasets, experimenting with different feature extraction techniques, and exploring alternative model architectures. By addressing these issues and continually iterating the model, the accuracy and reliability of emotion detection can be improved, ultimately resulting in a more robust and effective emotion sensing device.

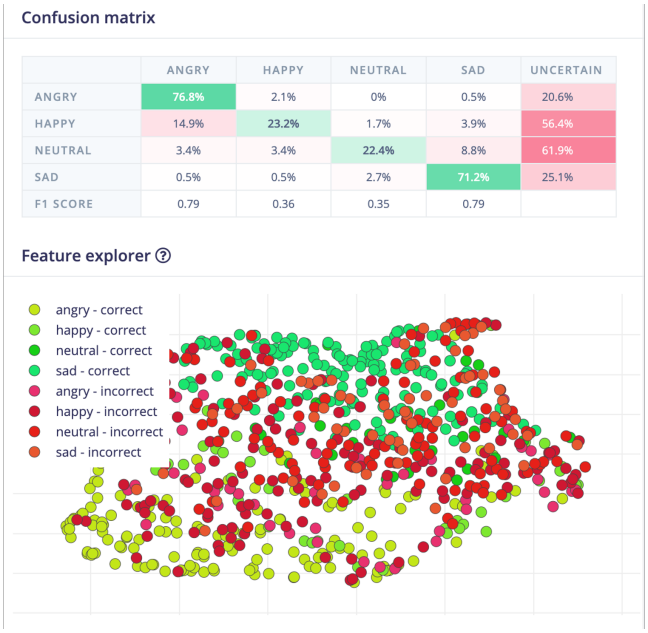

Figure 8: Result of Model Testing in Edge Impulse

4 Discussion and Future Work

The experimental evaluation and model testing conducted in Edge Impulse showed some concerning issues and limitations in the emotion-sensing device, suggesting that the current approach may not be optimal and that the audio data used for training could be problematic. One key issue that was identified was the potential presence of biases in the training data. The model demonstrated a tendency to associate high-volume audio clips with anger and low-volume clips with sadness, regardless of the actual emotional content. To address this issue, a possible solution would be to relabel the audio clips to include information about the speaker's gender, such as *female_happy*, or *male_angry*. By providing this additional context, the model could learn to recognize emotions more accurately across different types of voices.

Another area for improvement is the model's performance in detecting certain emotions, particularly happiness and neutral speech. To enhance the model's accuracy in these categories, it would be beneficial to collect and incorporate more training data specifically focused on these emotions. By exposing the model to a larger and more diverse set of examples, it can better learn the subtle differences and patterns associated with happiness and neutral speech, thereby improving its overall emotion recognition capabilities.

In addition to data collection efforts, trying different preprocessing techniques on audio clips can lead to significant improvements. Preprocessing involves applying various transformations and filters to the raw audio data, which is then fed into a model for training. By carefully selecting and optimizing these preprocessing steps, it is possible to highlight the features of the audio signal that are most relevant to the emotion, while suppressing noise and irrelevant information.

Looking beyond the current implementation, there is an exciting opportunity to extend the functionality of the emotion-sensing device by incorporating a music recommendation feature. By leveraging the detected emotions, the device could suggest personalized music selections to users based on their emotional state. For instance, if the device detects anger or stress in the user's voice, it could recommend calming and relaxing music to help alleviate those negative emotions. This feature would create a more engaging and interactive user experience, providing not only emotion detection but also a means to positively influence the user's emotional well-being.

To turn these ideas into reality, future work should focus on several key areas. First, the training dataset needs to be carefully extended, with particular emphasis on balancing the representation of different emotions and speaker demographics. Second, rigorous experiments and evaluations should be conducted to identify the most effective preprocessing techniques and model architectures for emotion detection. Finally, music recommendation features should be designed and integrated into the device taking into account user preferences, emotion mapping, and potential impact on user experience. By addressing these challenges and continuously refining emotion sensing devices, it has the potential to become a powerful tool for allowing machines to understand human emotions.

References

[1] American College Health Association. 2023. *American College Health Association-National College Health Assessment III: Undergraduate Student Reference Group Data Report Spring 2023.* American College Health Association, Silver Spring, MD. https://www.acha.org/documents/ncha/NCHA-III_SPRING_2023_UNDERGRAD_REFERENCE_GROUP_DATA_REPORT.pdf

[2] M. A. Venable and M. E. Pietrucha. 2022. *2022 College Student Mental Health Report.* BestColleges. https://healthymindsnetwork.org/wp-content/uploads/2023/08/HMS_National-Report-2022-2023_full.pdf

[3] M. A. Venable and M. E. Pietrucha. 2022. *2022 College Student Mental Health Report.* Retrieved June 14, 2024 from https://www.bestcolleges.com/wp-content/uploads/2023/06/Mental-Health_Report.pdf