PROJECT 6 REPORT

PRANIT SEHGAL

ASU ID : 1225456193

Voice Recognition Project Report on Arduino Nano 33 BLE Sense

***Project link**: https://studio.edgeimpulse.com/public/309235/latest*

A: Introduction

This project focused on creating a voice recognition system using the Arduino Nano 33 BLE Sense, with an aim to identify specific keywords associated with absolutist language. These keywords, relevant in the context of mental health diagnostics, include "all," "only," "must," "none," "never," and instances of silence. The system's goal was not to diagnose mental health conditions but to provide a potential tool for monitoring language markers that could be indicative of mental health states.

B: Experiment

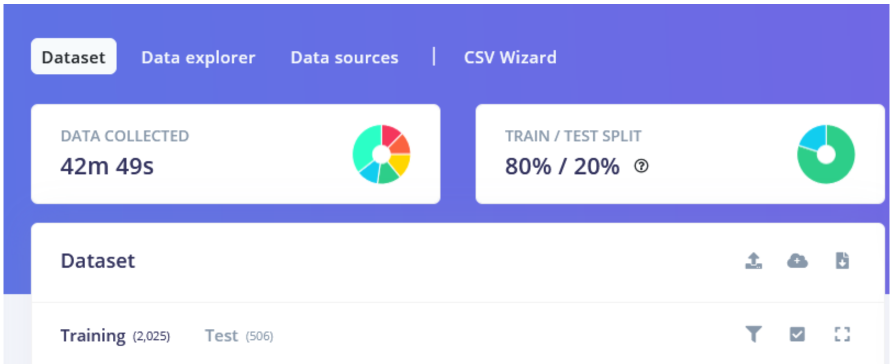
Data Collection and Dataset Construction

The audio dataset was meticulously gathered using two primary sources: my voice recordings and contributions from students in the class. Using the Harvard open speech recording tool, I recorded the words "all," "only," "must," "none," "never," and periods of silence. These recordings were stored in a dedicated folder named 'audio,' with separate subfolders for each keyword, all in .wav format.

This personalized dataset was augmented with the Simple Speech Command dataset to create a comprehensive collection for training the machine learning model.

Dataset Details

- **Number of Keywords**: 6 (including silence)
- **Data Sources**: Personal recordings and student contributions
- **Total Dataset Size**: 2531 samples (2025 for training, 506 for testing)
- **Dataset Split**: 80% training data (2025 samples), 20% testing data (506 samples)
- **Format**: .wav files
- **Labeling**: Each audio file was accurately labeled with its corresponding keyword or silence.



DATA COLLECTED
42m 49s

TRAIN / TEST SPLIT
80% / 20%

Dataset

Training (2,025) Test (506)

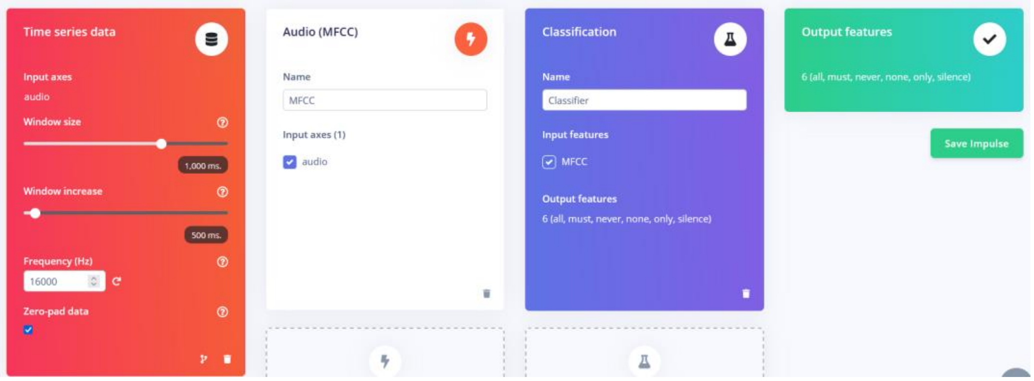
C: Algorithm

Design and Architecture

The machine learning model was designed using Edge Impulse software, emphasizing a window size of 1000ms compatible with MFCC processing. The model's architecture included an "Audio (MFCC) processing block" for extracting salient features from the audio data, and a basic classifier adept at recognizing patterns in audio.

Training Process and Parameters

- **Training Cycles**: 100
- **Learning Rate**: 0.005
- **Architecture**: Sequential 1D CNNs with a pooling layer (8 neurons, 3 kernel size)
- **Dropout Rate**: 0.25 for each layer
- **Flattening**: Applied to organize data effectively
- **Training Data**: 2025 samples
- **Testing Data**: 506 samples



Input axes
audio

Window size: 1,000 ms.

Window increase: 500 ms.

Frequency (Hz): 16000

Zero-pad data ☑

Name: MFCC

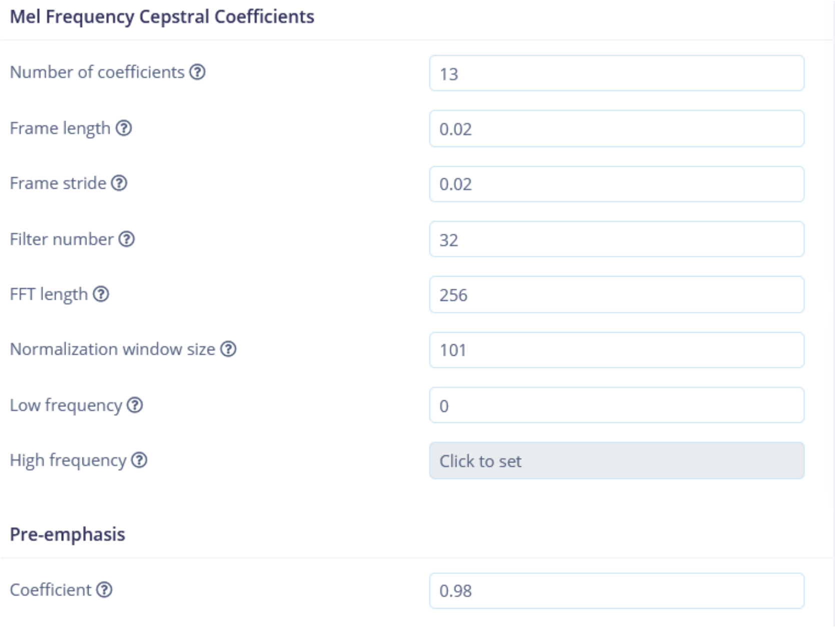
Input axes (1)
☑ audio

Name: Classifier

Input features
☑ MFCC

Output features
6 (all, must, never, none, only, silence)

6 (all, must, never, none, only, silence)

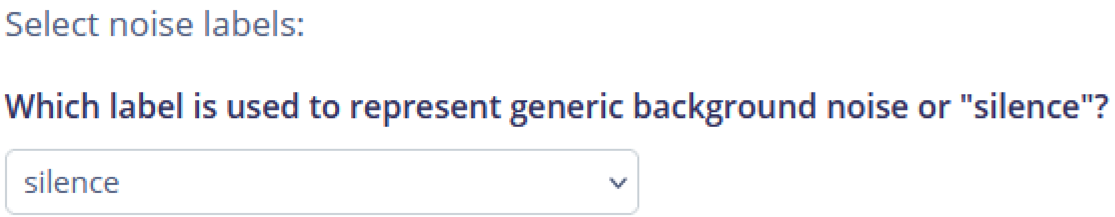


Mel Frequency Cepstral Coefficients

Number of coefficients	13
Frame length	0.02
Frame stride	0.02
Filter number	32
FFT length	256
Normalization window size	101
Low frequency	0
High frequency	Click to set

Pre-emphasis

Coefficient	0.98



Select noise labels:

Which label is used to represent generic background noise or "silence"?

silence

The training was successful, leading to a model that was both accurate and efficient in processing and recognizing the specified keywords.

D: Results

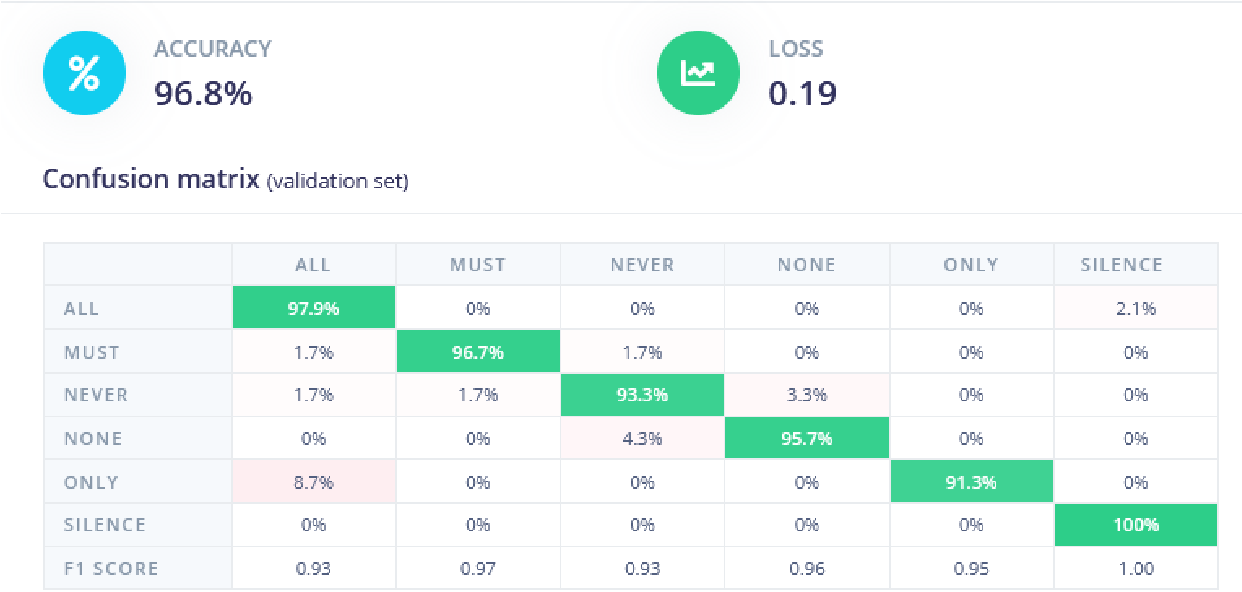
Model Performance

The model exhibited excellent performance metrics:

- **Accuracy**: 96.8%
- **Loss**: 0.19
- **Inference Time**: 3ms
- **Peak RAM Usage**: 3.8k
- **Flash Usage**: 31.6k

Confusion Matrix

The confusion matrix demonstrated the model's proficiency in identifying each keyword accurately:



ACCURACY

96.8%

LOSS

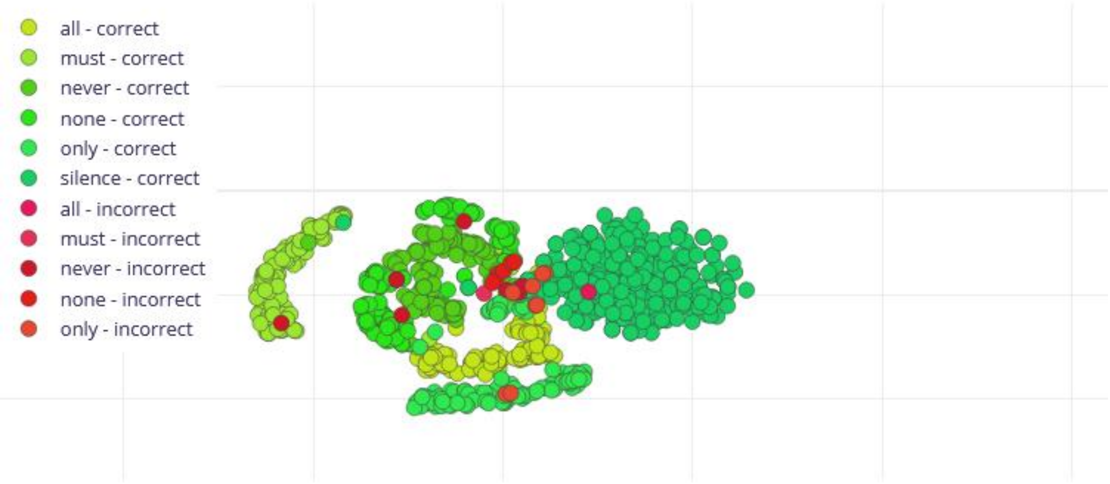
0.19

Confusion matrix (validation set)

	ALL	MUST	NEVER	NONE	ONLY	SILENCE
ALL	97.9%	0%	0%	0%	0%	2.1%
MUST	1.7%	96.7%	1.7%	0%	0%	0%
NEVER	1.7%	1.7%	93.3%	3.3%	0%	0%
NONE	0%	0%	4.3%	95.7%	0%	0%
ONLY	8.7%	0%	0%	0%	91.3%	0%
SILENCE	0%	0%	0%	0%	0%	100%
F1 SCORE	0.93	0.97	0.93	0.96	0.95	1.00

F1 Scores

- All: 0.93
- Must: 0.97
- Never: 0.93
- None: 0.96
- Only: 0.95
- Silence: 1.00



Real-time Prediction and Demonstration

Post-calibration, the model defaulted to "silence" in the absence of other keywords. Deployed as an Arduino header file, it was integrated into the Arduino system, demonstrating exceptional real-time performance and accuracy.

E: Discussions

Summary and Achievements

The project achieved its goal of developing a high-accuracy voice recognition system capable of identifying specific keywords. This system shows potential for use in continuous monitoring of mental health markers through speech.

Challenges and Future Directions

A significant challenge was ensuring the model's sensitivity to distinguish between similar sounding words and silence. Future work could explore more sophisticated audio processing techniques and neural network architectures to enhance accuracy and sensitivity.

Difficulties and Improvements

Integrating diverse audio data sources and training the model to recognize a fixed set of keywords was challenging. Future improvements could include expanding the dataset and experimenting with more advanced machine learning models for improved performance.

Accuracy of Real-time Predictions

The real-time predictions closely mirrored the expected accuracy, underscoring the model's reliability for practical applications. This capability suggests the system's suitability for real-world applications, particularly in wearable technologies for mental health monitoring.