Utilizing Liquid Neural Network for Efficient Audio Classification

Department of Electronic and Computer Engineering
Thapathali Campus

Team Members
Khemraj Shrestha (THA077BCT020)
Niyoj Oli (THA077BCT029)
Om Prakash Sharma (THA077BCT030)
Punam Shrestha (THA077BCT038)

Supervised By
Er. Kshetraphal Bohara

Co-Supervised By
Er. Rojesh Man Shikhrakar

July 2024

Presentation Outline

- Motivation
- Project Objectives
- Project Scope
- Project Applications
- Methodology
- Results
- Discussion of Results
- Remaining Tasks
- References

Motivation

- Contemporary models uses millions to billions parameters,
- 19 neurons enough for autonomous driving - Liquid Time Constant (LTC) Neural Network,
- Papers suggest LNN to be efficient for temporal data like Audio.

Project Objectives

- To develop LTC neural network model for audio classification and benchmark against contemporary models,
- To achieve comparable accuracy while using less computational power.

Project Scope

- Classify audio events with Liquid Neural Networks (LNN)
- Achieve high accuracy in real-time sound recognition
- Benchmark performance on multiple audio dataset for optimal results
- Revolutionize speech, environmental sound, and anomaly detection
- Optimize network architecture for diverse audio applications

Project Applications

- Voice Authentication and Security
- Medical Data Analysis
- Abnormality Detection
- Enhanced Music Recommendation Systems
- Audio Content Filtering and Moderation
- Interactive Gaming and Virtual Reality
- Speech Emotion Recognition
- Intelligent Audio Summarization

Liquid Neural Network-[1]
Introduction

- Traditional RNNs, face challenges in adapting to complex time-series dynamics.
- LNNs, still making use of recurrent mechanics, explicitly model time-series dynamics through differential equations that determine neuron states.

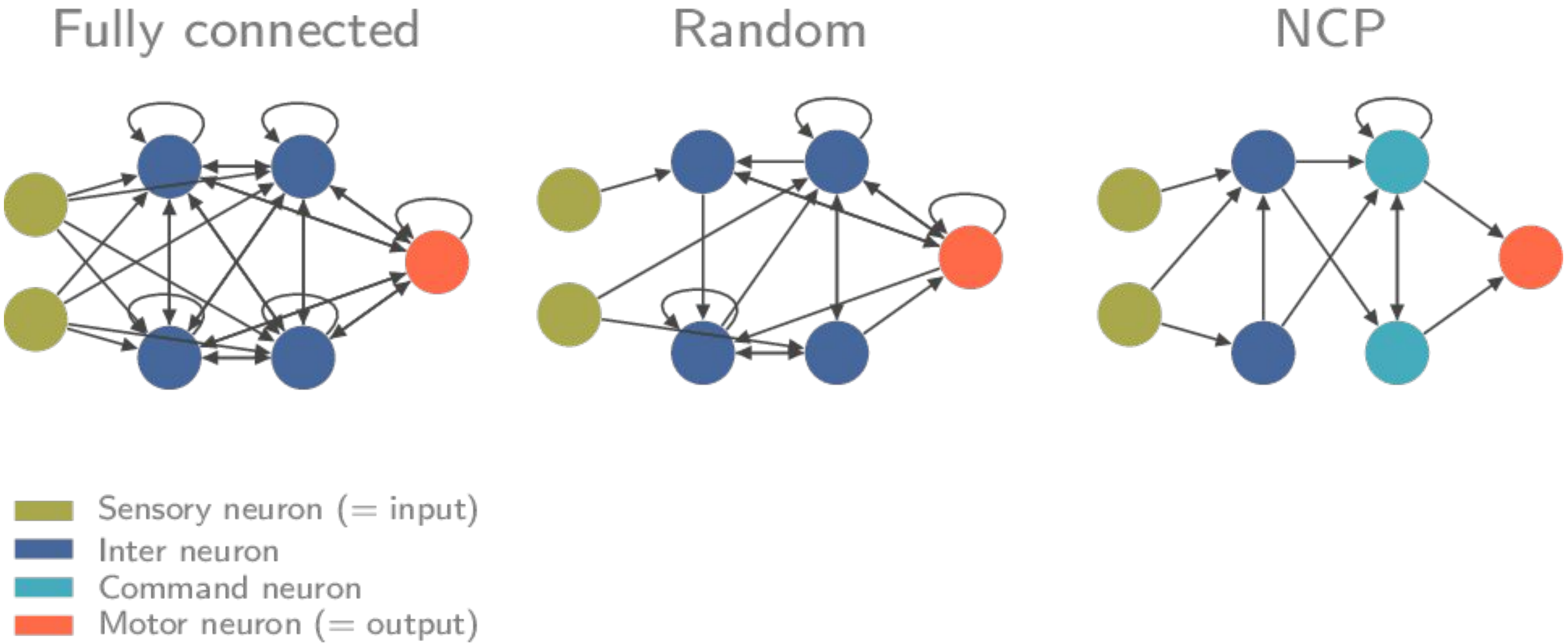
Liquid Neural Network-[2] Mathematical Formulation

- Neuron’s state is the solution to the differential equation

$$\frac{d\mathbf{x}(t)}{dt} = -\left[\frac{1}{\tau} + f(\mathbf{x}(t), \mathbf{I}(t), \boldsymbol{\theta})\right]\mathbf{x}(t) + f(\mathbf{x}(t), \mathbf{I}(t), \boldsymbol{\theta})\mathbf{A} \quad (1)$$

- Where,
  - *x(t)* is the hidden state
  - *I(t)* is the input
  - τ, time constant is a constant that ensures numerical stability

Liquid Neural Network-[3]
Neuron Structure



Liquid Neural Network-[4]
Wiring Configuration

Neuron Model

Liquid time-constant (LTC)

Closed-form continuous-time (CfC)

Wiring

Fully Connected

```
from ncps.torch import LTC
input_size = 20
units = 50
rnn = LTC(input_size, units)
```

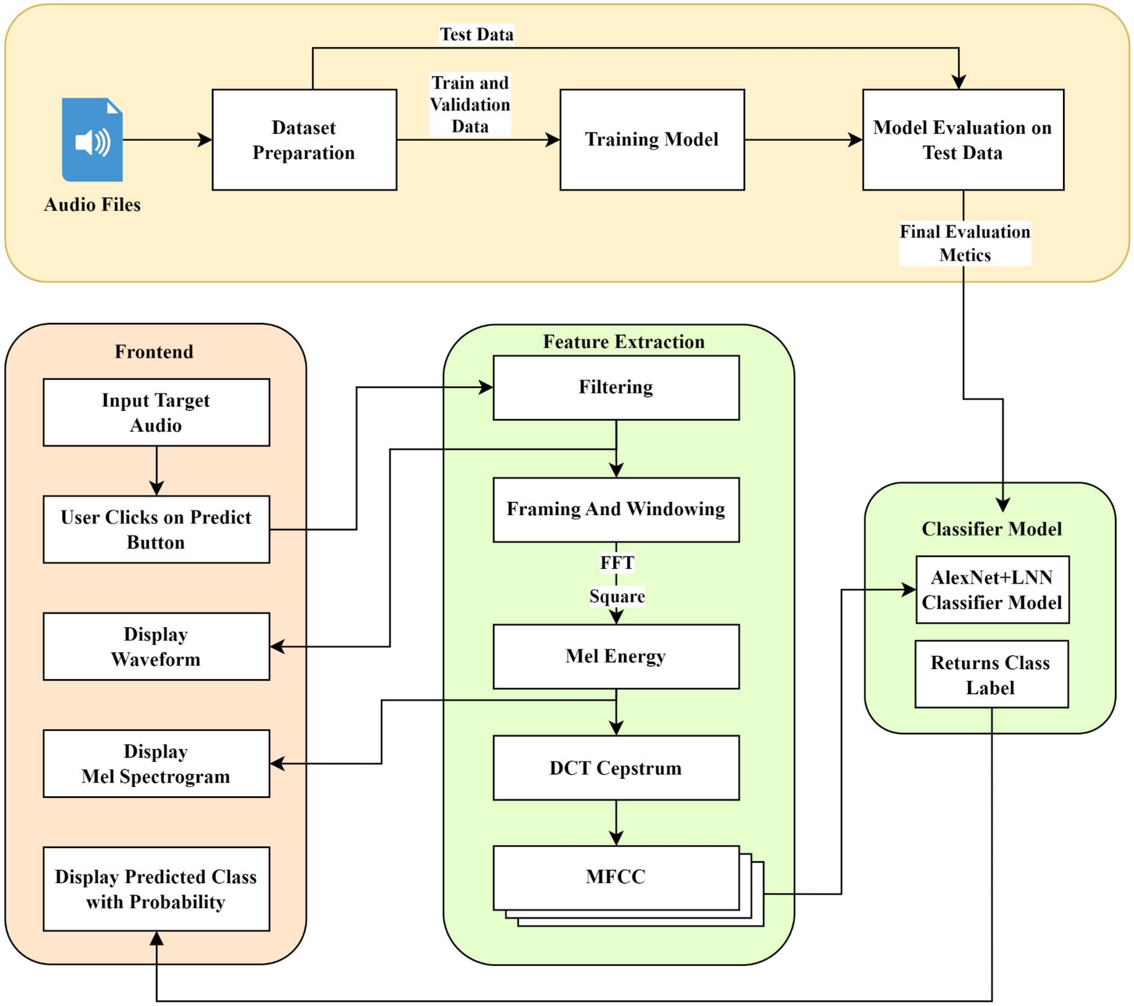
```
from ncps.torch import CfC
input_size = 20
units = 50
rnn = CfC(input_size, units)
```

Neural Circuit Policy (NCP)

```
from ncps.torch import LTC
from ncps.wirings import AutoNCP
input_size = 20
units = 50
output_size = 10
rnn = LTC(input_size,
    AutoNCP(units,output_size))
```

```
from ncps.torch import CfC
from ncps.wirings import AutoNCP
input_size = 20
units = 50
output_size = 10
rnn = CfC(input_size,
    AutoNCP(units, output_size))
```

Methodology-[1]
System Architecture



Methodology-[2]
Dataset Exploration

1. **VGG (Visual Geometry Group)**

- Audio-visual dataset with 210,000 data with 310 audio classes,
- Classes like wind noise, sliding door, car, train, etc.
- 10-second audio clips,
- Roughly 200 audio per each class,
- Used by Mirasol3B has 69.8% accuracy on this dataset.

Methodology-[3]
Dataset Exploration

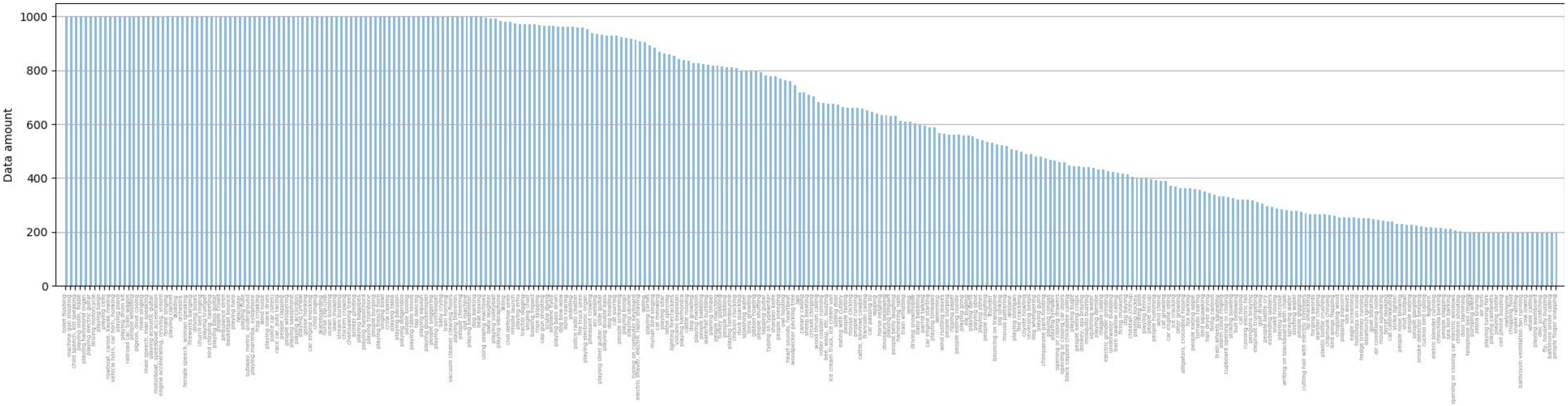


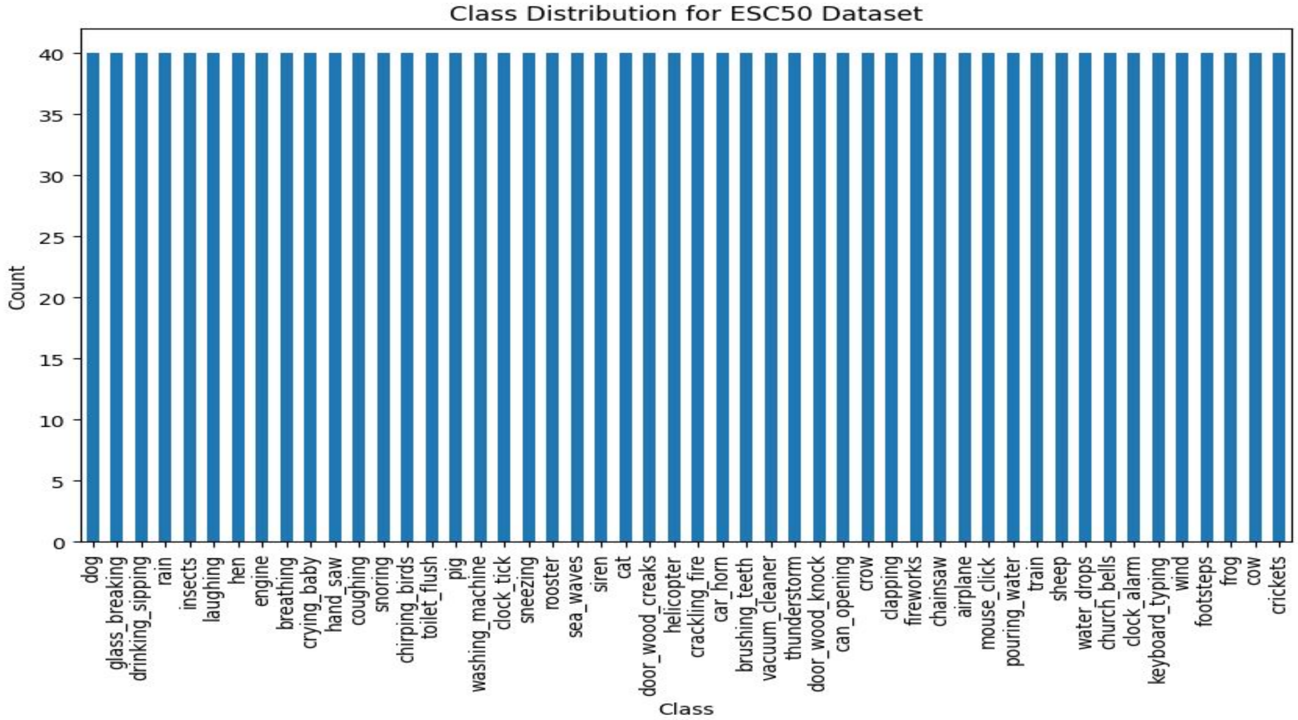
Fig: Dataset distribution for VGG Dataset

Methodology-[4]
Dataset Exploration

2. ESC-50

- 2000 labelled environmental audio recordings,
- Each clip of 5 seconds, covering 50 distinct classes,
- Includes classes like animals, water sound, natural soundscapes, etc.
- Pre-arranged in 5-folds,
- Used by OmniVec-2 Model with 99.1% accuracy.

Methodology-[5]
Dataset Exploration

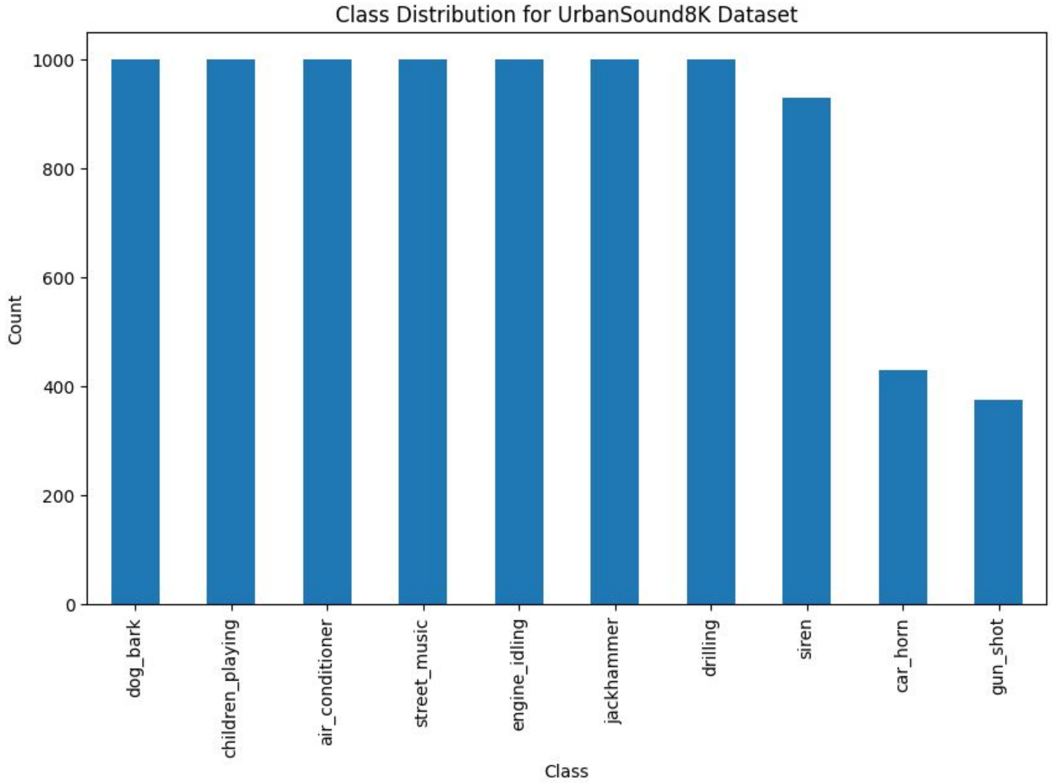


Methodology-[6]
Dataset Exploration

3. UrbanSound8K

- Comprising 8,732 labelled sound excerpts,
- Each clip of 4 seconds, with total 27 hours of audio,
- Includes classes like air conditioner, car horn, children playing, dog bark, etc,
- Used by ASM-RH-I with 97.96% accuracy (10-fold).

Methodology-[7]
Dataset Exploration

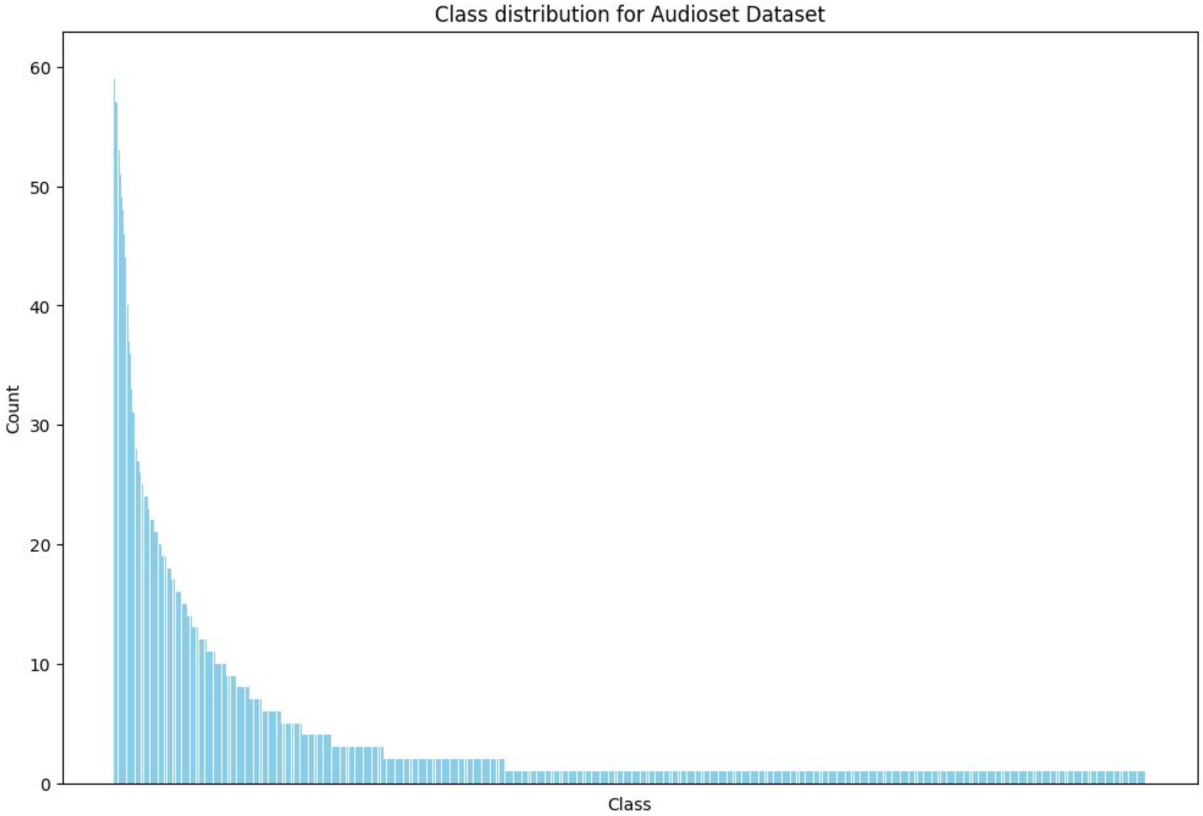


Methodology-[8]
Dataset Exploration

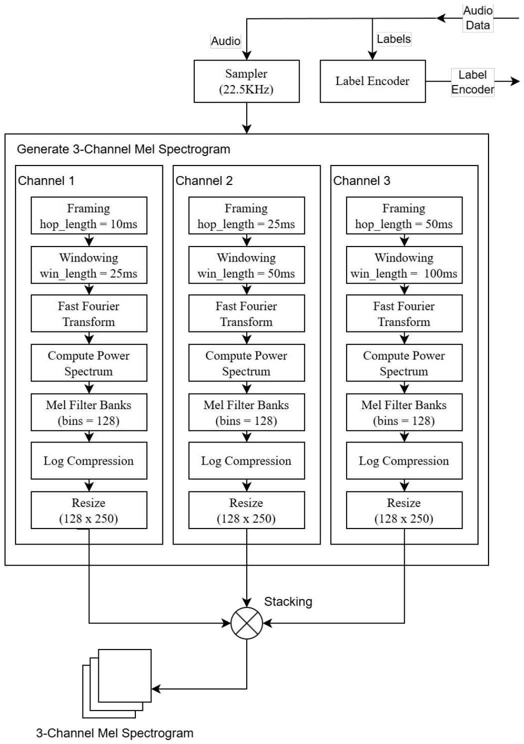
4. AudioSet

- 2,084,320 YouTube videos containing 527 labels,
- 10-second sound clips sourced from YouTube videos and labelled by humans,
- Includes classes like music, speech, vehicle, car, etc.,
- Used by OmniVec with 0.548 mAP.

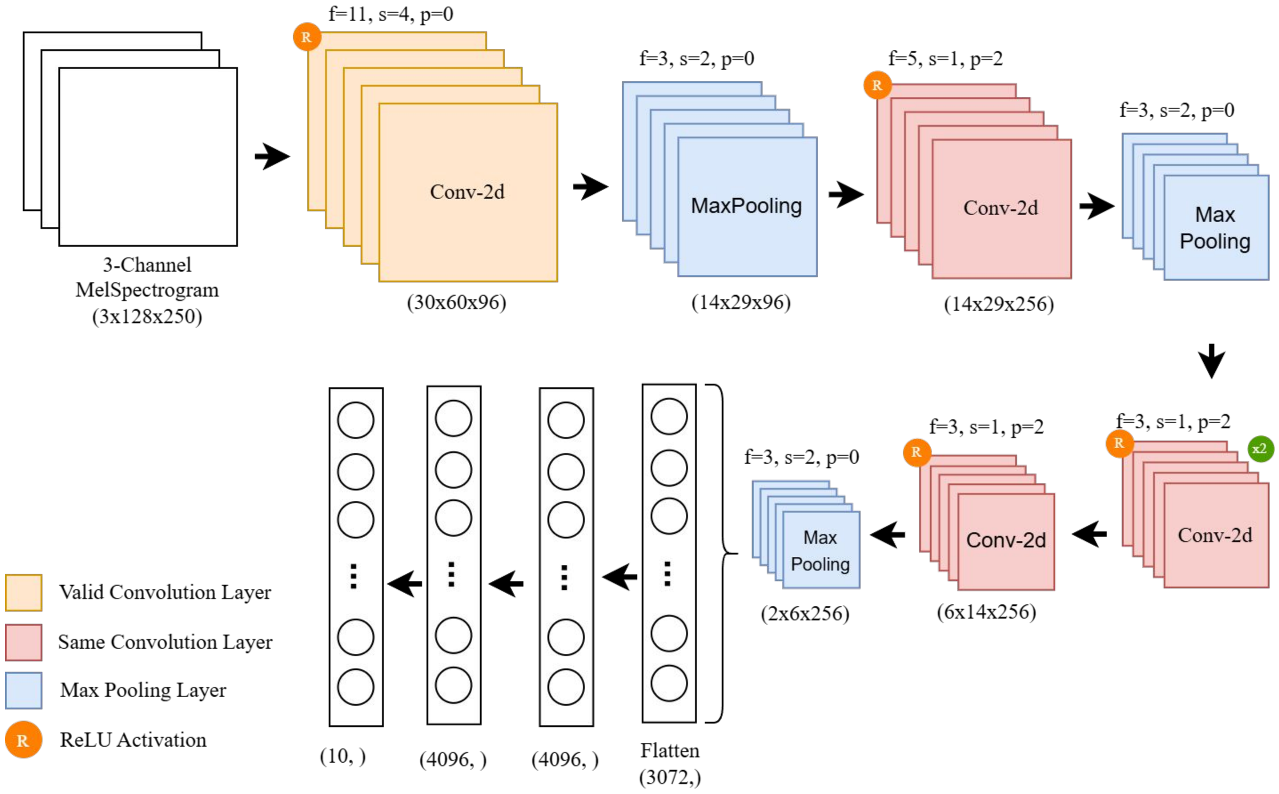
Methodology-[9]
Dataset Exploration



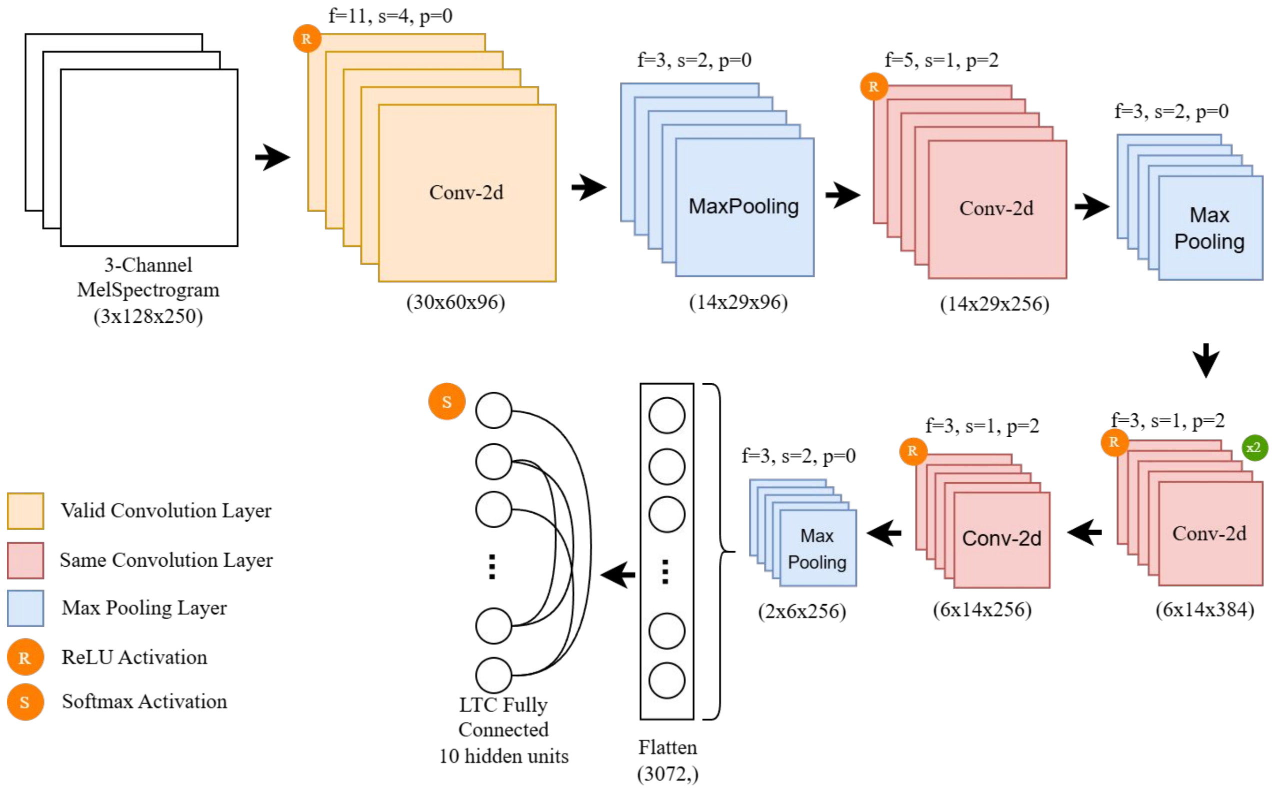
Methodology-[10]
Data PreProcessing



Methodology-[11]
Baseline Model (CNN)



Methodology-[12]
Our Model (CNN + LTC)



Methodology-[13]
Evaluation Metrics

- **F1-Score**

- Used when the class distribution is imbalanced,
- Provides a single measure that balances both the false positives and false negatives.
- **Precision** is the ratio of true positive detections to the total number of positive detections,
- **Recall** is the ratio of true positive detections to the total number of actual positives.

Methodology-[14]
Evaluation Metrics

- **F1-Score**

- Harmonic mean of precision and recall.

$$F1Score = \frac{2 \cdot Recall \cdot Precision}{Recall + Precision}$$

Methodology-[15]
Evaluation Metrics

- **Mean Average Precision (mAP)**

- **Average Precision** is the area under the precision-recall curve for a single query or class.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i$$

Methodology-[16]
Evaluation Metrics

- **Accuracy**

- measures the proportion of correct predictions made by the model out of all predictions.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{All Predictions}}$$

Methodology-[17]
Instrumentation

1. **Kaggle Notebook**

- Kaggle Notebooks are essentially Jupyter Notebooks hosted on the cloud
- Provides 4 CPU cores, 20GB of RAM, and 1 x Nvidia Tesla P100 GPU with 4 cores and 29 GB of RAM,
- GPU can be used for 30 hours a week and 9 hours per session.

Methodology-[18]
Instrumentation

2. Google Colaboratory

- Provides an Intel Xeon CPU with 2 vCPUs (virtual CPUs) and 13 GB of RAM,
- NVIDIA Tesla K80 with 12GB of VRAM (Video Random-Access Memory)

Methodology-[19]
Instrumentation

3. Librosa

- Python package for music and audio analysis,
- Calculation of time domain features like Zero-crossing rate,
- Calculation of frequency domain features.

Methodology-[20]
Instrumentation

4. Pytorch

- Open-source deep learning framework developed by Facebook's AI Research lab,
- Uses Dynamic Computation Graph,
- Rich ecosystem and community support.

Results - [1]

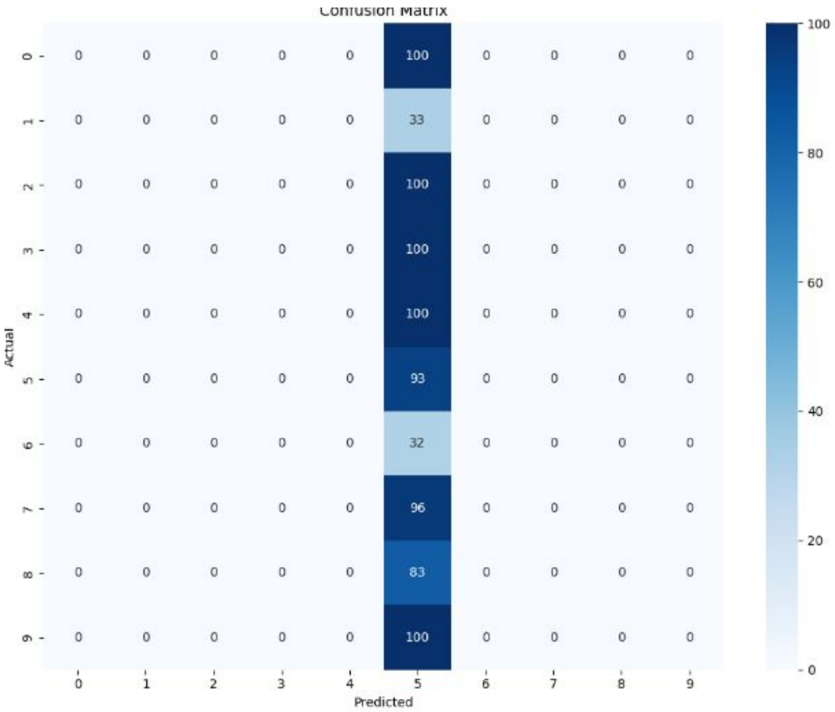


Fig: Confusion Matrix

Results - [2]

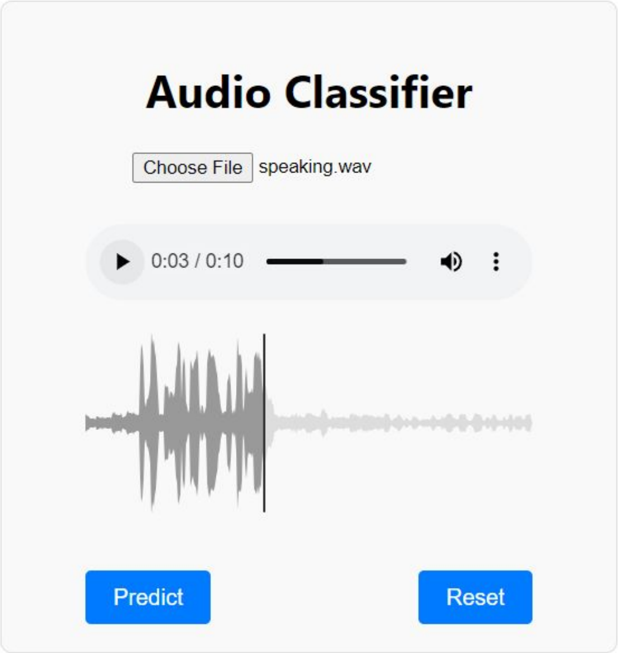


Fig: Web application interface

Results - [3]

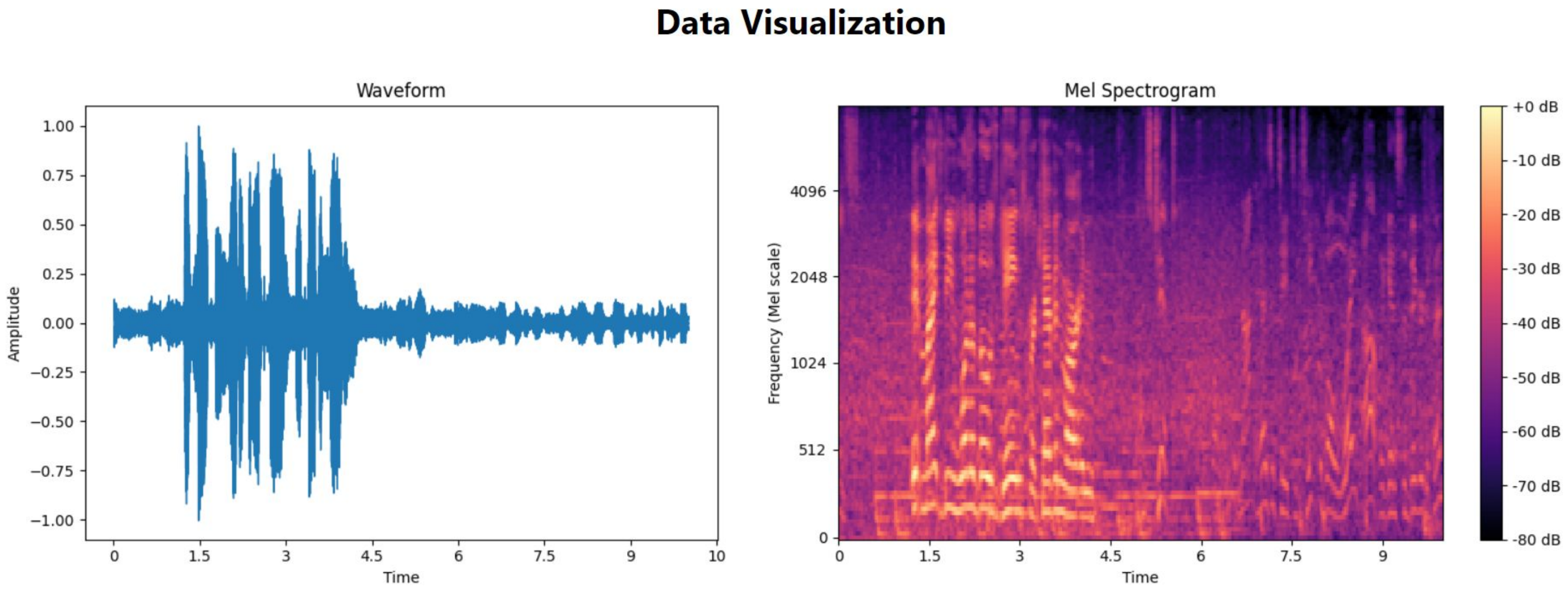


Fig: Web application interface

Results - [4]

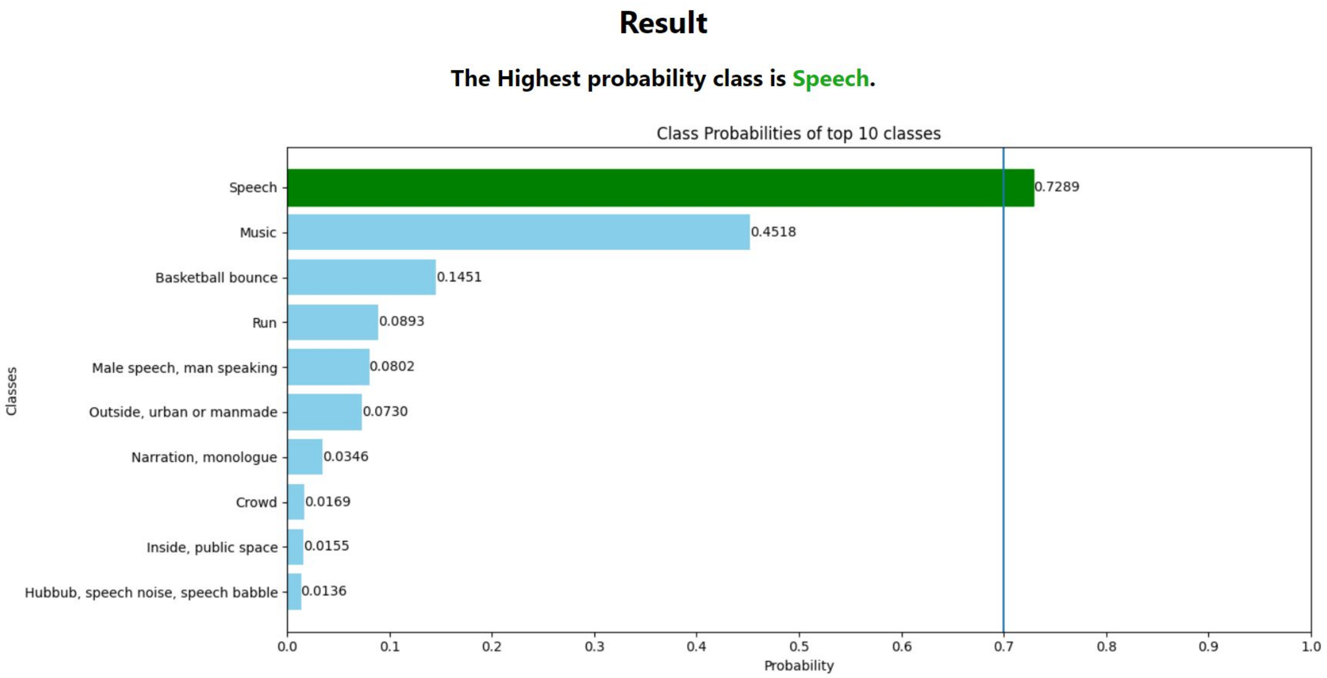


Fig: Web application interface

Discussion of Results - [1]

Frontend:

- User-Friendly Interface for Seamless Audio Interaction
- Upload .wav or .mp3 Files for Instant Audio Spectrograms
- Predict Top 10 Audio Classes Quickly and Accurately
- Single File Upload with Cancel Option for Flexibility

Discussion of Results - [2]

Backend:

- **Model Integration:** Uses a LNN with CNN architecture for audio classification.
- **Feature Extraction:** Extracts and normalizes features like mel-spectrograms from uploaded audio.
- **Visualization:** Generates and serves waveform plots, bar plots and mel-spectrograms.
- **Error Handling:** Manages file upload errors and provides user-friendly error messages.
- **Adaptive Precision and Hardware Utilization:** Uses Automatic Mixed Precision (AMP) for faster computation on GPUs, and switches to CPU if a GPU is unavailable.

Remaining Tasks

- The models need to be tuned and modified for better accuracy,
- Different configuration of LTCs is to be tried,
- Methods like knowledge distillations are needed to be explored for decreasing model size.

Reference - [1]

[1] R. Hasani, M. Lechner, A. Amini, D. Rus, and R. Grosu, “Liquid time-constant networks,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 9, 7657–7666, May 2021. DOI: 10.1609/aaai.v35i9.16936. https://ojs.aaai. org/index.php/AAAI/article/view/16936.

[2] S. Srivastava and G. Sharma, Omnivec: Learning robust representations with cross modal sharing, 2023. arXiv: 2311.05709 [cs.CV].

Reference - [2]

[3] M. Chahine, R. Hasani, P. Kao, et al., “Robust flight navigation out of distri bution with liquid neural networks,” Science Robotics, vol. 8, no. 77, eadc8892, 2023, Published online 2023 Apr 19, ISSN: 2470-9476.

[4] H. Ju, J.-X. Xu, and A. M. VanDongen, “Classification of musical styles using liquid state machines,” in The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, 1–7.