SPEECH
SPEECH
SPEECH
CLASSIFICATION

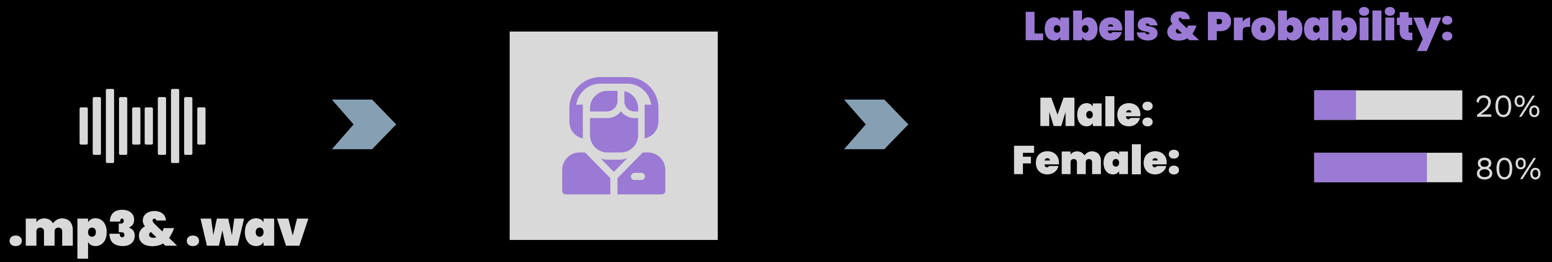
AUDIO ENGINEER PORTFOLIO AUDIO ENGINEER PORTFOLIO

GROUP MEMBER: Harish Ram, Zeqiu Zhang, Jiachen Sands, Sisi Zhang

TABLE OF CONTENTS

PROJECT MOTIVATION

- Understand how speech data can be utilized to understand emotion, race, age, sex, accent and Parkinson’s Disease
- Our goal is to develop a highly-accurate, pre-trained Speech Engine
- Designed a GUI to show the classification results based on user input



Speech Recognition Tasks

EMOTIONS	AGE	RACE	GENDER	ACCENT	PARKINSON'S DISEASE
6 Labels: Sad, Fear, Happy, etc	5 Labels: <30, 30s, 40s, 50s, etc	3 Labels: African American, Caucasian, Asian	2 Labels: Male, Female	5 Labels: English, French, Mandarin, etc	2 Labels: Negative, Positive

DATA

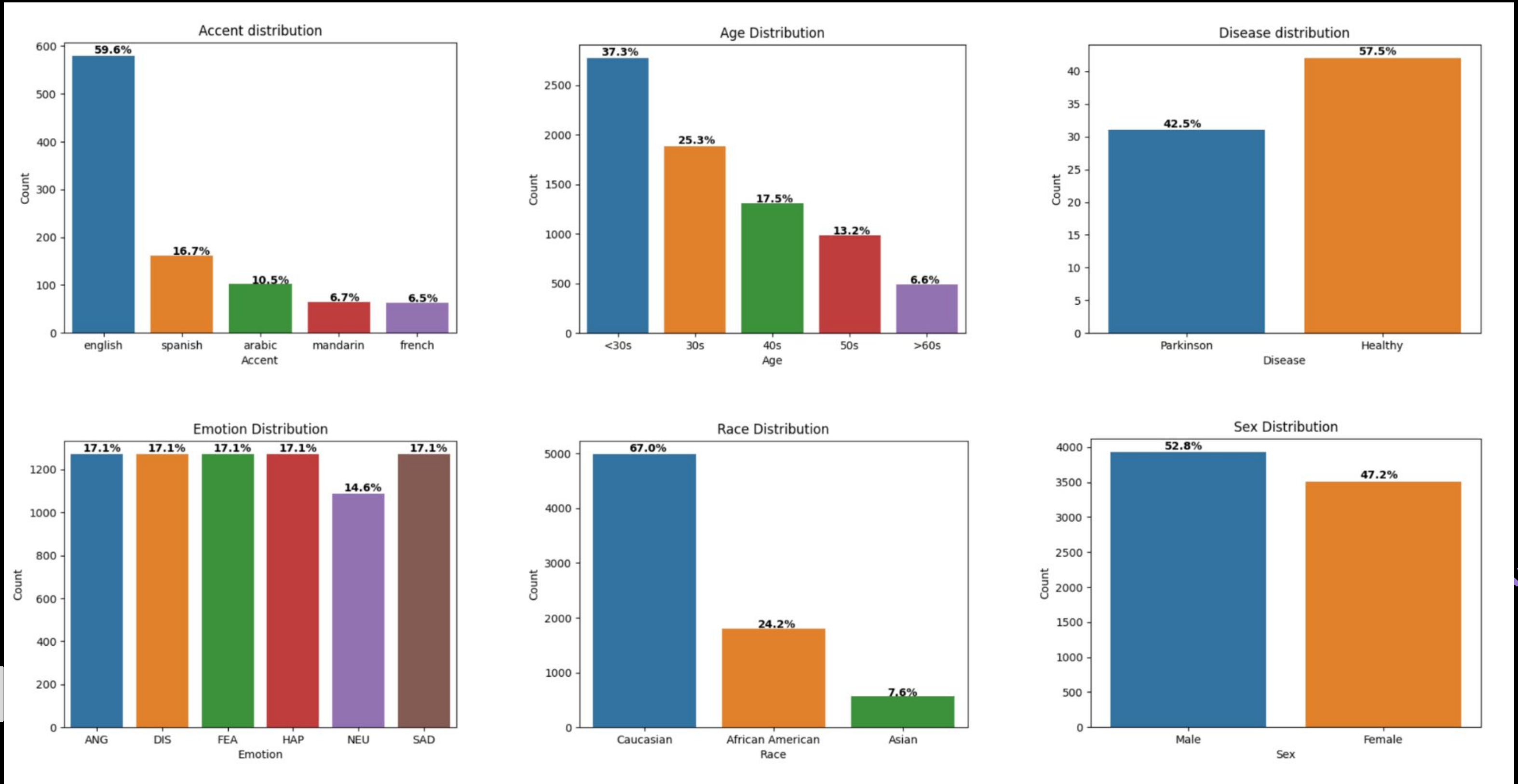
4 Data Sources:

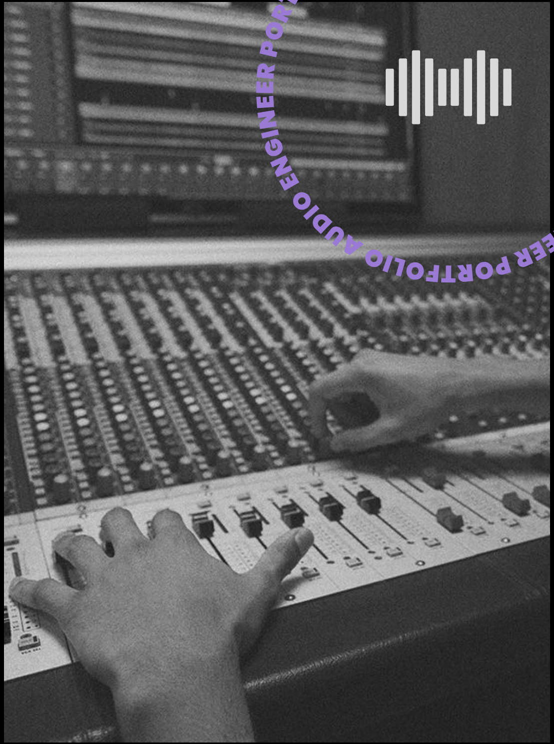
- Emotions, Race, Age, Sex: 7443 .wav files (CREMA-D)
- Accent: 971 .mp3 files for Top 5 labels (Kaggle)
- Parkinson’s Disease: 73 .wav files (MDVR-KCL)
- Pseudo-Labeling For Parkinson’s Disease: LJSpeech (~13,100 short clips)

Issues:

- Imbalanced Datasets (Age, Accent, Race)
- Large Audio Files (Accent, Parkinson’s Disease)

DATA DISTRIBUTION





DATA SOLUTIONS

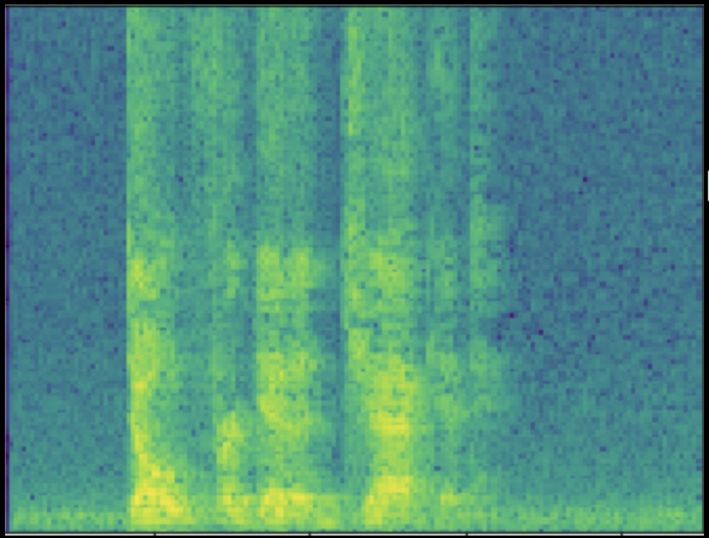
- Split Large Audio Files & Remove Silence
  - Generated 5566 More .wav Files For Accent
  - Generated 657 More .wav Files for Parkinson's Disease
- Combine Prediction Results:
  - Accent: Mode
  - Parkinson's Disease: Average
- Data Augmentation Methods
  - Adding White Noise
  - Time Shift
  - Pitch Scale, etc
- Autoencoder
  - Condensing and restructuring Mel Spectrogram images for feature extraction
- Pseudo-Labeling
  - Adding unlabeled data to increase sample size of training set

OLIO AUDIO ENGINEER PORTF

MEL SPECTROGRAM

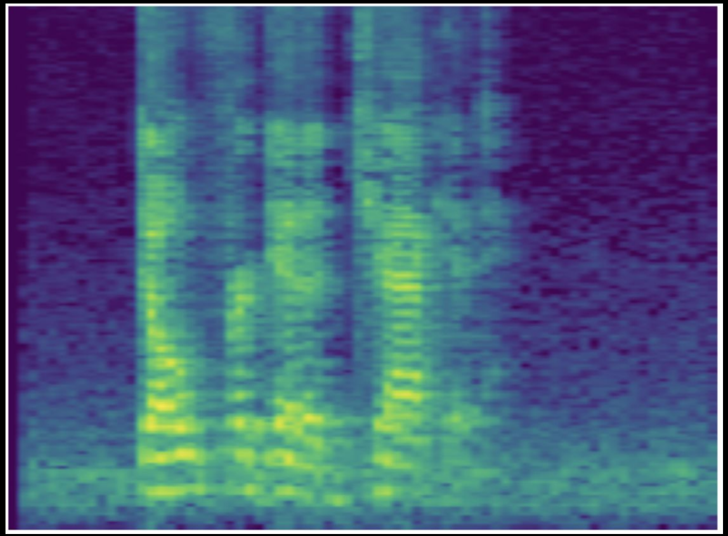
SPECTROGRAM

Frequency



Time

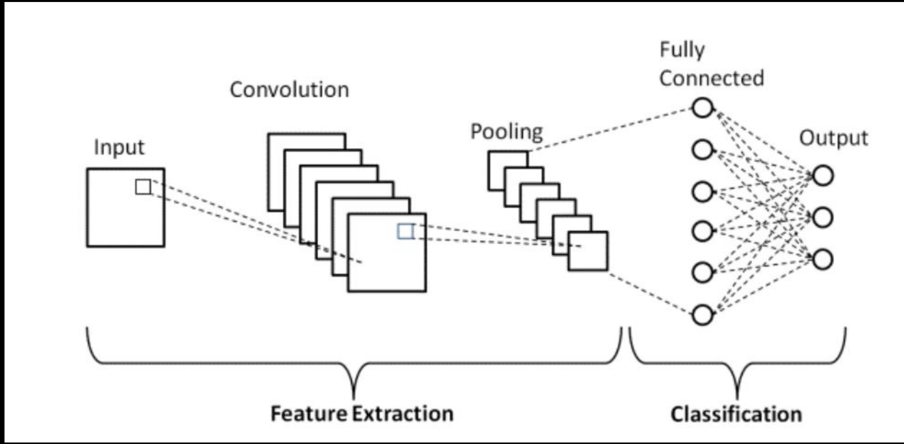
MEL SPECTROGRAM



N_ftt: 1024
Hop_length: 512
n_mels:128

- It uses the Mel Scale instead of Frequency on the y-axis
- It uses the Decibel Scale instead of Amplitude to indicate colors

CNN



Convolutional

- Kernel and Stride
- Matrix Multiplication on Image
- Feature Extraction

Pooling

- Max Pooling
- Reduce Computation Power

Fully Connected

- Linear transformation
- Flatten and returns single vector with class probabilities

Activation

- Computationally Efficient
- Applying gradient calculation
- ReLU

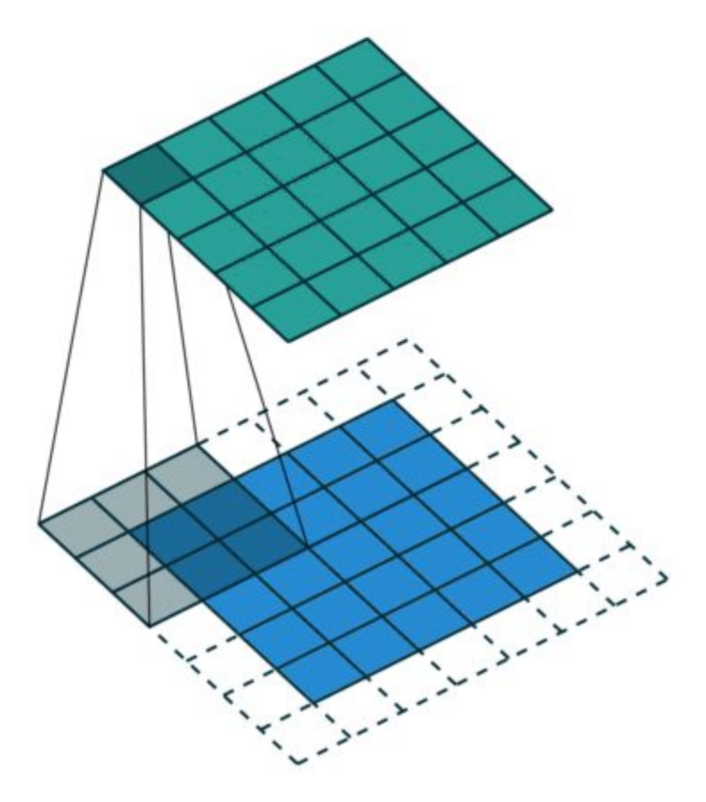
BENCHMARK

Benchmark - CNN3			
Category	#Label	Accuracy	F1
Accent*	5	58.97%	58.97%
Age*	5	52.79%	51.79%
Disease	2	74.47%	69.07%
Emotion	6	46.88%	46.43%
Race*	3	72.62%	72.16%
Sex	2	92.21%	92.21%
Note: * indicates class imbalance			

PRETRAINED MODELS

There were several convolutional pre-trained models that we used for calculate benchmark scores:

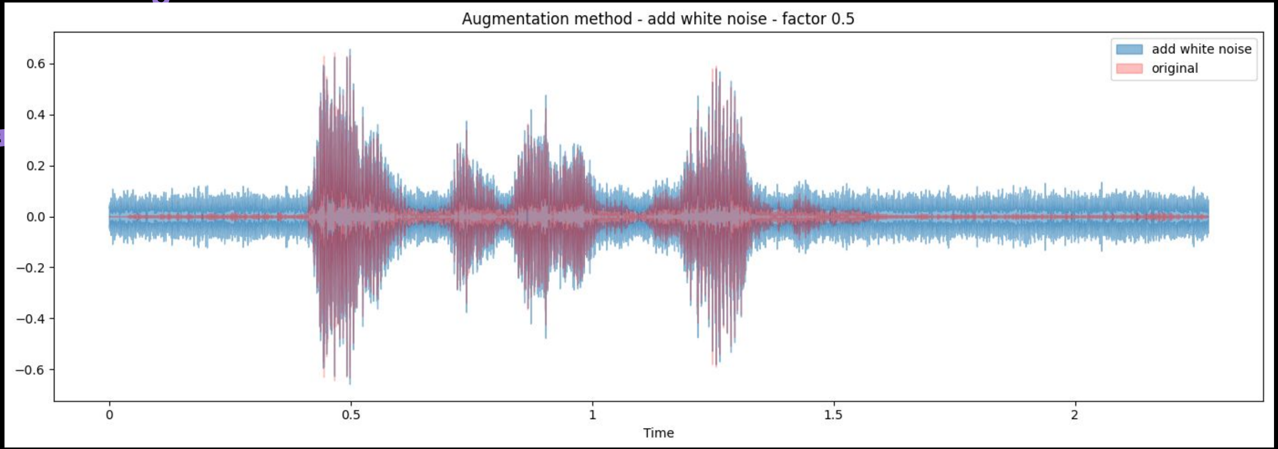
- Resnet18
- Resnet34
- VGG16
- EfficientNet_b2



BEST CNN MODEL

Best CNN Model by Category				
Category	**#Label**	**Model**	**Accuracy**	**F1**
Accent*	5	ResNet34	61.10%	60.51%
Age*	5	ResNet18	82.87%	82.73%
Disease	2	CNN9	93.19%	91.21%
Emotion	6	CNN9	54.67%	54.43%
Race*	3	ResNet34	89.95%	89.75%
Sex	2	ResNet18	97.85%	97.85%
Note: * indicates class imbalance				

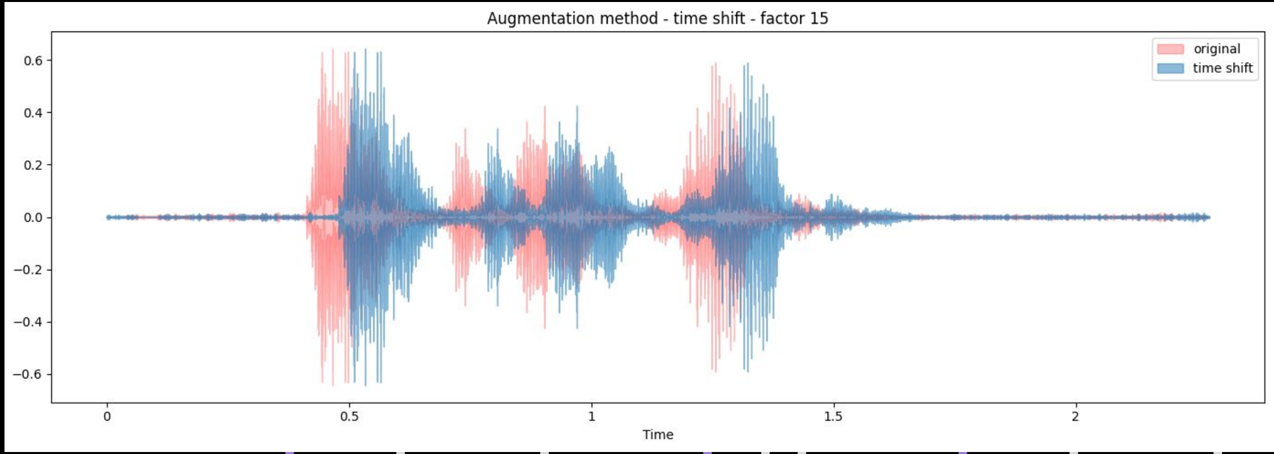
AUGMENTATION



- 1st Graph: Add white noise with factor 0.5
- 2nd Graph: Time shift with factor 15

Implemented the following augmentation methods randomly with random factors:

- Add white noise
- Time shift
- Time stretch

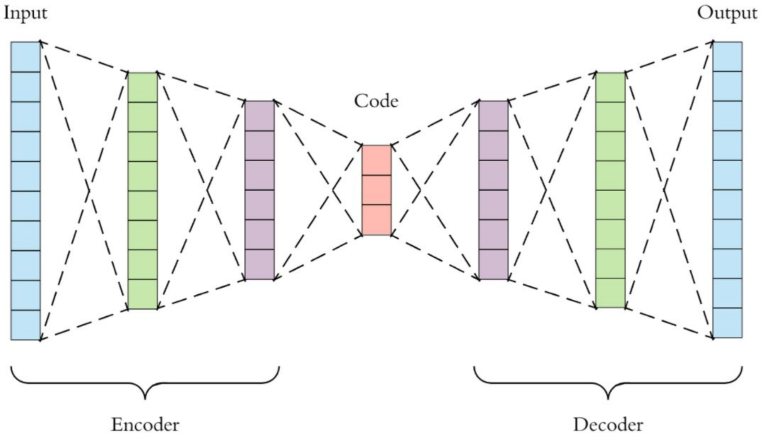


AUGMENTATION RESULT

Model F1 Score Before and After Augmentation				
Category	**Model**	**Pre-Augmentation F1 Score**	**Augmentation Method**	**Post-Augmentation F1 Score**
Accent	ResNet34	60.51%	Time Shift, White Noise	59.71%
Age	ResNet18	82.73%	Time Stretch, White Noise	80.92%
Race	ResNet34	89.75%	Time Shift	90.23%

- Combined augmented data with original data (doubled sample size)
- Applied combined data on the best Pre-trained models

AUTOENCODER



Encoder

- Image is compressed
- Representation of Image is generated

Decoder

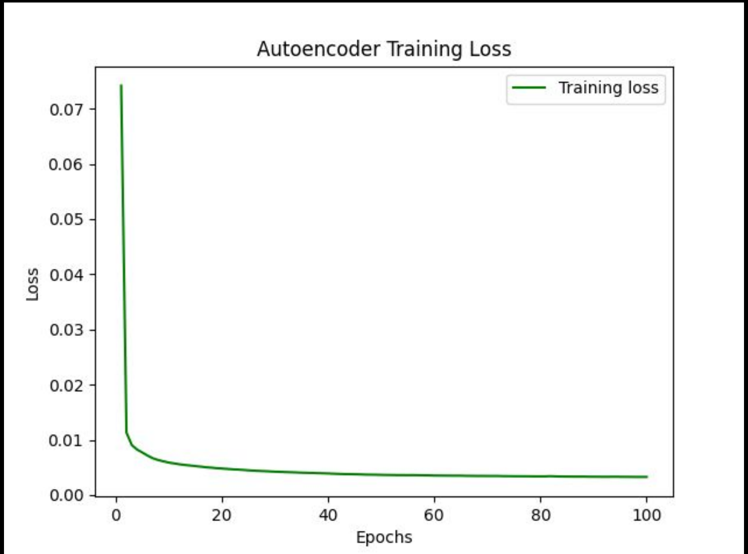
- Reconstruct Image with same size as input

Purpose

- Pre-training for CNN/Pre-trained models
- Use weights and biases as starting point

AUDIO ENGINEER PORTFOLIO

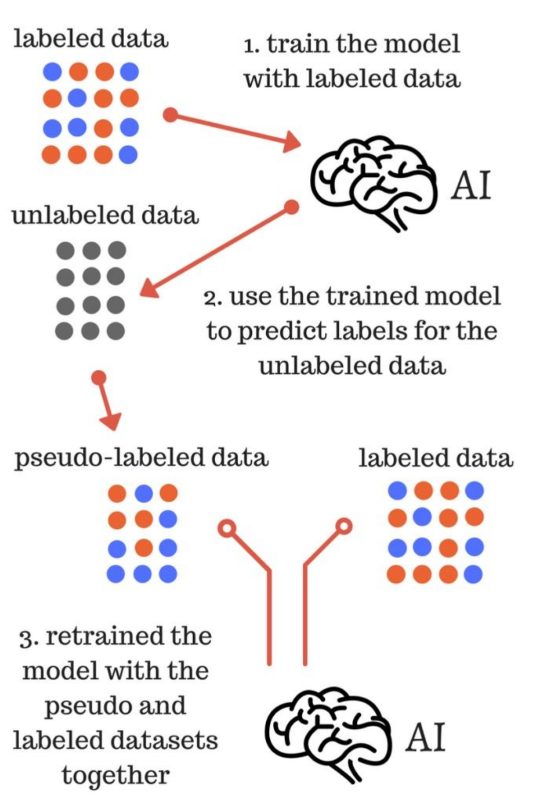
AUTOENCODER GRAPH



Using Autoencoder for Pre-training			
Category	**#Label**	**CNN3 F1**	**CNN3 (with AE) F1**
Accent*	5	58.97%	57.43%
Age*	5	51.79%	54.78%
Disease	2	69.07%	72.25%
Emotion	6	46.43%	47.50%
Race*	3	72.16%	75.82%
Sex	2	92.21%	94.36%

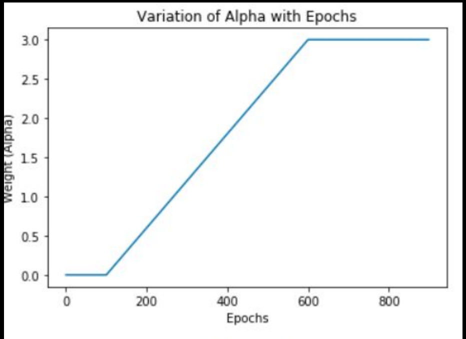
Note: * indicates class imbalance

PSEUDO-LABELING



Loss = Labeled Loss + Alpha * Unlabeled Loss

$$\alpha(t) = \begin{cases} 0 & t < T_1 \\ \dfrac{t - T_1}{T_2 - T_1}\alpha_f & T_1 \leq t < T_2 \\ \alpha_f & t \geq T_2 \end{cases}$$



Variation of Alpha with Epochs

PSEUDO-LABELING RESULTS

Using Pseudo-Labeling on Best CNN Model				
Category	**#Label**	**Model**	**Before F1**	**After F1**
Accent*	5	ResNet34	60.51%	58.37%
Age*	5	ResNet18	82.73%	88.82%
Disease	2	CNN9	91.21%	75.13%
Emotion	6	CNN9	54.43%	54.21%
Race*	3	ResNet34	89.75%	89.28%
Sex	2	ResNet18	97.85%	98.05%

Note: * indicates class imbalance

PORTFOLIO AUDIO ENGINEER

FINAL SCORES

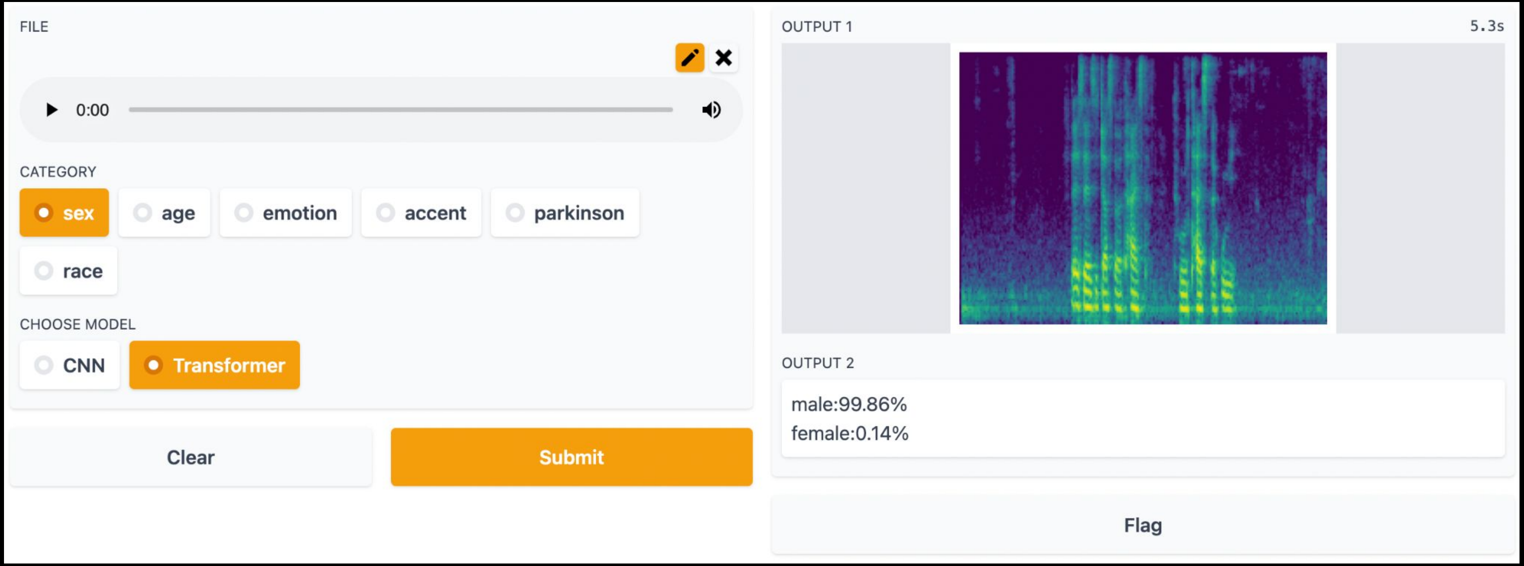
Speech Classification F1 Scores							
Categories	#Label	CNN3 - Benchmark	CNN3 (with AE)	Best CNN*	Best CNN* (with Aug.)	Best CNN* (with PL)	Wav2Vec2-base
Accent*	5	58.97%	57.43%	60.51%	59.71%	58.37%	63.07%
Age*	5	51.79%	54.78%	82.73%	80.92%	88.82%	85.30%
Disease	2	69.07%	72.25%	91.21%	---	75.13%	94.67%
Emotion	6	46.43%	47.50%	54.43%	---	54.21%	76.05%
Race*	3	72.16%	75.82%	89.75%	90.23%	89.28%	94.63%
Sex	2	92.21%	94.36%	97.85%	---	98.05%	99.40%

Note: * indicates class imbalance; Best CNN* including CNN9, ResNet18 and ResNet34.

GUI

GRADIO LIBRARY

- Fully connect all models of all categories into GUI
- User is able to choose CNN models or Wav2Vec2 Model



REFERENCE

[1] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458*.

[2] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into imaging*, *9*(4), 611-629.

[3] Dertat, A. (2017, October 8). *Applied deep learning - part 3: Autoencoders*. Medium. Retrieved March 21, 2022, from https://towardsdatascience.com/applied-deep-learning-part-3-autoencoders-1c083af4d798

[4] Shenoy, A. (2019, December 3). *Pseudo-labeling to deal with small datasets-what, why & how?* Medium. Retrieved April 1, 2022, from https://towardsdatascience.com/pseudo-labeling-to-deal-with-small-datasets-what-why-how-fd6f903213f

[5] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449-12460.

[6] Wei, S., Zou, S., & Liao, F. (2020). A comparison on data augmentation methods based on deep learning for audio classification. In *Journal of Physics: Conference Series* (Vol. 1453, No. 1, p. 012085). IOP Publishing.

[7] Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, *72*(2011), 1-19.

[8] Lee, D. H. (2013, June). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML* (Vol. 3, No. 2, p. 896).

[9] Li, Z., Ko, B., & Choi, H. J. (2019). Naive semi-supervised deep learning using pseudo-label. *Peer-to-peer networking and applications*, *12*(5), 1358-1368.

[10] Wikimedia Foundation. (2022, April 28). *Convolutional Neural Network*. Wikipedia. Retrieved April 29, 2022, from https://en.wikipedia.org/wiki/Convolutional_neural_network

PORTFOLIO AUDIO ENGINEER

REFERENCE

[11] Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017, August). Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)* (pp. 1-6). Ieee.

[12] Deng, L. (2012). The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, *29*(6), 141-142.

[13] Velardo, V. (2022, January 1). *Musikalkemist/audiodataaugmentationtutorial: Repository hosting code and slides of the audio data augmentation series on the sound of ai yt channel.* GitHub. Retrieved February 2, 2022, from https://github.com/musikalkemist/audioDataAugmentationTutorial

[14] pytorch.org. (n.d.). *Conv2d*. Conv2d - PyTorch 1.11.0 documentation. Retrieved March 22, 2022, from https://pytorch.org/docs/stable/generated/torch.nn.Conv2d.html

[15] pytorch.org. (n.d.). *ConvTranspose2d*. ConvTranspose2d - PyTorch 1.11.0 documentation. Retrieved March 22, 2022, from https://pytorch.org/docs/stable/generated/torch.nn.ConvTranspose2d.html

[16] Admin. (2021, November 16). *Python remove silence in WAV using Librosa - Librosa tutorial*. Tutorial Example. Retrieved February 14, 2022, from https://www.tutorialexample.com/python-remove-silence-in-wav-using-librosa-librosa-tutorial/

[17]Eracube. (2020, February 7). *Split audio on timestamps librosa*. Stack Overflow. Retrieved February 14, 2022, from https://stackoverflow.com/questions/60105626/split-audio-on-timestamps-librosa

[18] Rachael Tatman (2018). *Speech Accent Archive*. Data Retrieved from https://www.kaggle.com/datasets/rtatman/speech-accent-archive?select=recordings

[19] Cao H, Cooper DG, Keutmann MK, Gur RC, Nenkova A, Verma R. *CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset*. *IEEE transactions on affective computing. 2014;5(4):377-390*. doi:10.1109/TAFFC.2014.2336244.

[20] Hagen Jaeger, Dhaval Trivedi, & Michael Stadtschnitzer. (2019). *Mobile Device Voice Recordings at King's College London (MDVR-KCL) from both early and advanced Parkinson's disease patients and healthy controls [Data set]*. Data retrieved from https://doi.org/10.5281/zenodo.2867216

[21] Ito, K., & Johnson, L. (2017). *The LJ speech dataset*. The LJ Speech Dataset. Retrieved April 2022, from https://keithito.com/LJ-Speech-Dataset/

[22] Kodžoman, V. (2019, May 19). Pseudo-labeling a simple semi-supervised learning method. Retrieved April 29, 2022, from https://datawhatnow.com/pseudo-labeling-semi-supervised-learning/

OLIO AUDIO ENGINEER PORTFO

THANK YOU