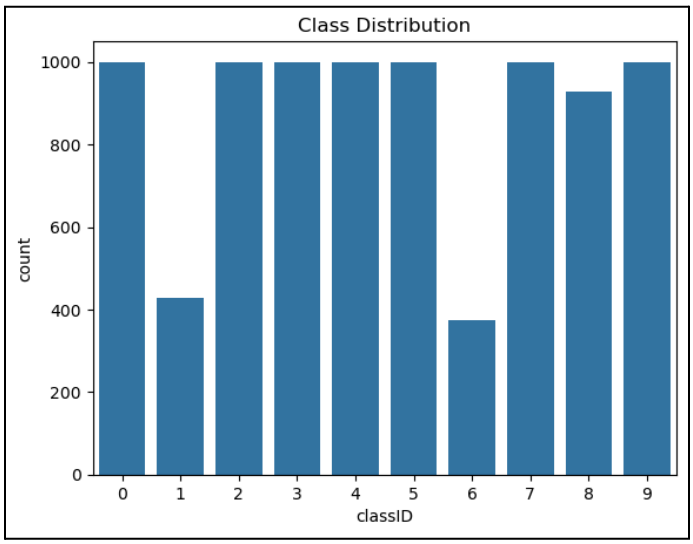
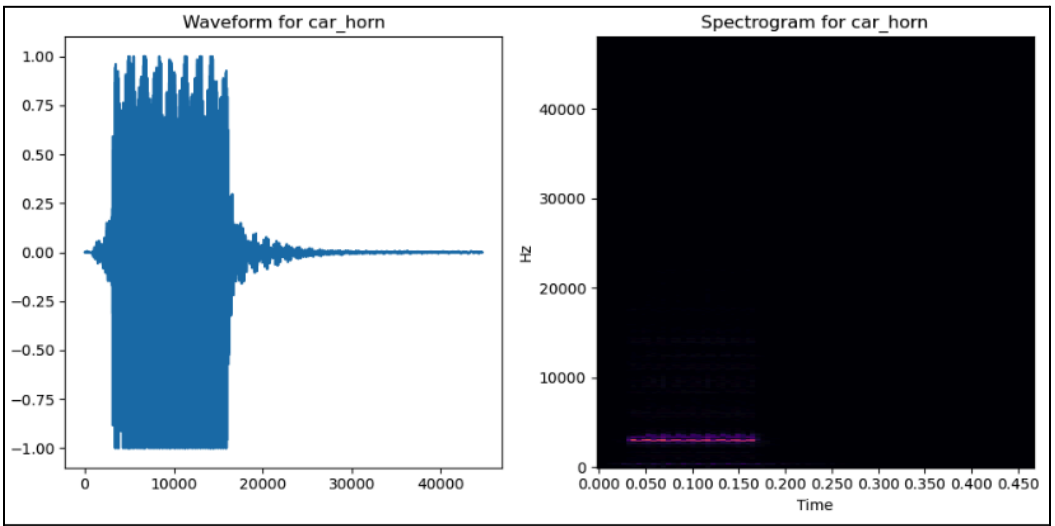
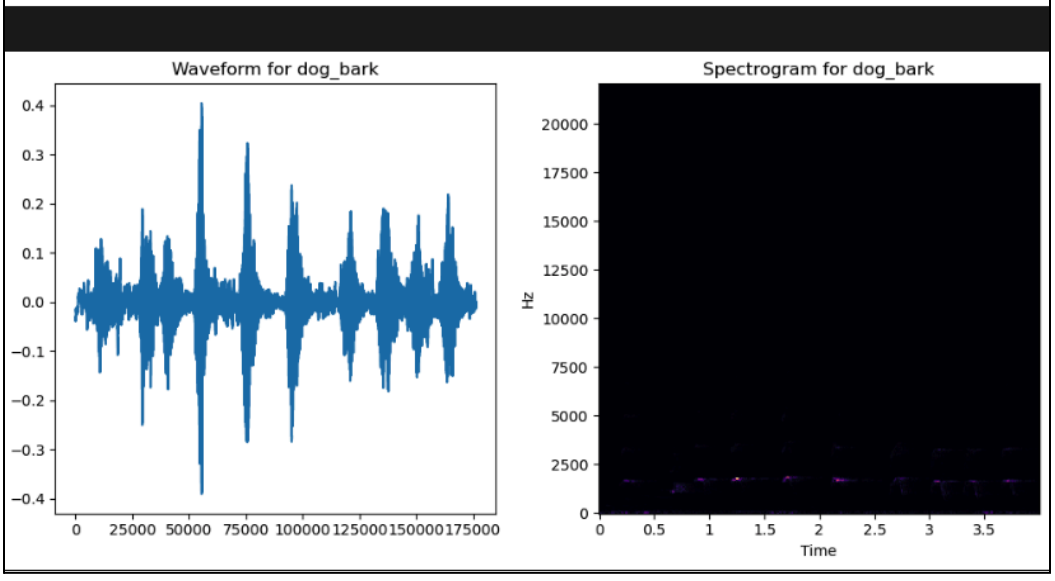
Introduction :

This report summarizes the methodologies and techniques implemented in the speech classification task. The task aimed to classify audio signals into predefined categories using state-of-the-art deep learning and machine learning methods. Several approaches were explored, their results analyzed, and future directions proposed for improving the model's performance.

Data Exploration :







Approaches Applied :

1. **Using MFCC features for classification (Training Various Classifiers on Extracted features)**

MFCCs were extracted to summarize spectral properties of audio signals. Other augmentations like noise addition and pitch shifting were implemented for data diversification.

These features were directly fed into initial experiments, achieving reasonable results but showing limitations in capturing variations in the data.

	Classifier	Accuracy	Precision	Recall
7	XGBClassifier	90.097310	90.696887	88.922086
2	RandomForestClassifier	89.639382	91.291343	87.747162
0	KNeighborsClassifier	88.036634	88.535161	87.292217
6	MLPClassifier	87.406983	88.101102	87.108376
4	GradientBoostingClassifier	80.995993	82.468859	79.779925
1	DecisionTreeClassifier	69.147109	68.131359	68.035459
5	SGDClassifier	52.718947	51.598587	50.852382
3	AdaBoostClassifier	40.297653	40.163358	41.739581

2. **Using Wav2Vec2 Embeddings for Classification (Training Various Classifiers on extracted embeddings)**

Used Wav2vec2 embeddings for the audio signals and used them to train the following classifiers and obtained the given result :

- **Metrics Included:**
  - Accuracy, Precision, Recall, and F1 Score.
  - Confusion Matrix to visualize misclassifications.

	Classifier	Accuracy	Precision	Recall
6	MLPClassifier	76.817401	76.988950	76.278934
7	XGBClassifier	73.325701	74.737083	72.405966
2	RandomForestClassifier	67.830567	69.978907	66.794056
0	KNeighborsClassifier	67.429880	70.309213	67.352268
5	SGDClassifier	67.315398	70.274456	67.254525
4	GradientBoostingClassifier	64.796795	66.434941	64.332101
1	DecisionTreeClassifier	46.651402	46.305037	46.358933
3	AdaBoostClassifier	34.459073	37.307939	35.413271

Data Augmentation

- **Methodology:**
  - Techniques like noise injection, time-stretching, pitch shifting, and mixing multiple audio sources were explored.
  - Augmented data was combined with the original dataset for training.
- **Outcome:** The augmented data improved validation accuracy slightly, suggesting the model's sensitivity to overfitting on limited original data.

3. ZSC and Fine Tuning Wav2Vec2 Model for Classification

Why Used: Wav2Vec2, a pretrained model from Facebook, is designed to process raw waveforms and extract high-level features. It has been widely used for speech-to-text tasks, but this project repurposed it for classification.

- **Methodology:**
  - Fine-tuned the Wav2Vec2 model directly on the classification task with additional linear layers for output logits.
  - Experimented with both the full fine-tuning approach and a parameter-efficient LoRA (Low-Rank Adaptation) setup to determine optimal configurations.
  - Loss functions like CrossEntropyLoss were employed for categorical classification.
- **Outcome:** Fine-tuning Wav2Vec2 achieved moderate results with high computational cost. The LoRA-based approach reduced the number of trainable parameters significantly but did not outperform full fine-tuning.

Optimizers, Learning Rate Schedulers, Dropouts,

- AdamW was chosen for its ability to handle sparse gradients effectively.

- ReduceLROnPlateau was used to adapt the learning rate based on validation loss, ensuring the model could focus on difficult learning phases.

Due to limited GPU resources and the given time I could only train the model for 1 epoch so I compared the 1 epoch trained and Zero Shot classification performance of Wav2Vec2

1. **Pretrained Model Performance:** Wav2Vec2 demonstrated promising capabilities by leveraging its pretrained contextual representations. However, its performance was limited due to:
    - Insufficient fine-tuning data.
    - Task-specific adaptations were not optimal for classification.
2. **Challenges in Fine-Tuning:**
    - Large models like Wav2Vec2 require significant computational resources and data to generalize well.
    - Overfitting was evident in the validation loss trends, requiring advanced regularization techniques.

	Classifier	Accuracy	Precision	Recall
0	Wav2Vec2 + LoRA	0.195192	0.069279	0.161142

```
  _warn_prf(average, modifier, msg_start, len(result))
     Metric  Zero-Shot Model  Fine-Tuned Model
0  Accuracy          0.195192          0.195192
1 Precision          0.069279          0.069279
2    Recall          0.161142          0.161142
3  F1-Score          0.075213          0.075213
```

Future Directions :

Based on the findings, the following approaches are proposed to enhance performance:

1. **Advanced Data Augmentation:**
    - Use techniques like SpecAugment, which modifies spectrograms directly by masking time or frequency bands.
    - Simulate real-world environments by adding background noises (cafeteria noise, traffic, etc.).
2. **Transfer Learning with Domain-Specific Models:**
    - Fine-tune Wav2Vec2 models trained on domain-specific datasets.
    - Explore lightweight models like DistilWav2Vec for efficiency.
3. **Balanced Dataset Creation:**

    - Address class imbalance by oversampling underrepresented classes or generating synthetic data.
4. **Ensemble Learning:** **(*Reference 1 in Research Thesis)***
    - **Combine predictions from multiple models (e.g., CNN, RNN, and Wav2Vec2).**
    - **Use weighted averaging to improve robustness.**
5. **Architectural Changes:**
    - Experiment with hybrid architectures combining CNN for feature extraction and attention-based mechanisms for temporal understanding.
    - Introduce regularization layers like dropout and batch normalization to prevent overfitting.
6. **Hyperparameter Tuning:**
    - Use grid or random search to optimize parameters like learning rate, dropout, and batch size

.

I was in the process of trying :
Reference 1 and Thesis Research Reference 1

For Approach 1 - Refer MFCC.ipynb
For Approach 2 - Refer wav2vec.ipynb

Thesis paper Reference:

1. Reiser, L., & Fivian, A. (2021). Speech classification using deep learning. Zurich University of Applied Sciences (ZHAW). https://www.zhaw.ch/storage/engineering/institute-zentren/cai/BA21_Speech_Classification_Reiser_Fivian.pdf

References :

1. Rezaul, K. M., Jewel, M., Islam, M. S., Siddiquee, K. N. e. A., Barua, N., Rahman, M. A., Shan-A-Khuda, M., Sulaiman, R. B., Shaikh, M. S. I., Hamim, M. A., Tanmoy, F. M., Haque, A. U., Nipun, M. S., Dorudian, N., Kareem, A., Farid, A. K., Mubarak, A., Jannat, T., & Asha, U. F. T. (2024). Enhancing audio classification through MFCC feature extraction and data augmentation with CNN and RNN models. International Journal of Advanced Computer Science and Applications, 15(7), 37–53. https://doi.org/10.14569/IJACSA.2024.0150704
2. Vaessen, N., & van Leeuwen, D. A. (2021). Fine-tuning wav2vec2 for speaker recognition. arXiv. https://arxiv.org/abs/2109.15053

Articles of Help :

1. Neurotech Africa. (2020, October 2). Audio classification using deep learning. Medium. https://medium.com/neurotech-africa/audio-classification-using-deep-learning-a6585292c055
2. Chowdhury, A. (2018, December 4). Music genre classification using deep learning: Audio and video. Medium. https://medium.com/@aritrachowdhury95/music-genre-classification-using-deep-learning-audio-and-video-770173980104